**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**
**«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ**
**імені ІГОРЯ СІКОРСЬКОГО»**
**Інститут прикладного системного аналізу**
**Кафедра математичних методів системного аналізу**

«До захисту допущено»
В.О. завідувача кафедри
_____ О.Л. Тимощук

# Дипломна робота

**на здобуття ступеня бакалавра**

**з напряму підготовки 6.050101 "Комп'ютерні науки"**

**на тему: «Моделі, методи та короткострокове прогнозування демографічних процесів в Україні»**

Виконав:
студент IV курсу, групи КА-55
Тертичний Роман Віталійович
_____

Керівник:
професор кафедри ММСА,
д.т.н. Бідюк П. І. _____

Консультант з економічного розділу:
доцент, к.е.н. Шевчук О. А. _____

Консультант з нормоконтролю:
доцент, к.т.н. Коваленко А.Є. _____
Рецензент:
к.т.н. Мурга М.О _____

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ – 2019 року

# РЕФЕРАТ

Дипломна робота: 108 с., 24 рис., 8 табл., 2 додатки, 15 джерел.

ДЕМОГРАФІЧНІ ПРОЦЕСИ УКРАЇНИ, РЕГРЕСІЙНІ МОДЕЛІ, ПРОГНОЗУВАННЯ, ЧИСЕЛЬНІСТЬ НАСЕЛЕННЯ, НАРОДЖУВАНІСТЬ, СМЕРТНІСТЬ

Дана робота присвячена вивченню методів та моделей прогнозування, а саме авторегресійних моделей різних типів та порядків, розглянутих в практичній області демографічних процесах в Україні.

Метою дипломної роботи є дослідження поведінки часових рядів та їх побудова на основі авторегресійних моделей різних типів, а також розробка програмного продукту для отримання практичних результатів та вибору найкращої моделі для наочної візуалізації даних короткострокового прогнозування демографічних процесів народонаселення України.

Об’єктом дослідження є статистичні дані процесів народжуваності, смертності та загальної популяції населення України конвертовані в часові ряди.

# ABSTRACT

The work consist of 108 pages, 24 images, 8 tables, 2 append., 15sources.

The theme: «Models and methods for short-term forecasting of demographic processes in Ukraine».

DEMOGRAPHIC PROCESSES OF UKRAINE, REGRESSION MODELS, PREDICTION, TOTAL POPULATION, BIRTH RATE, MORTALITY RATE.

This work is devoted to the study of methods and models of forecasting, namely autoregressive models of different types and orders, considered in the practical field of demographic processes in Ukraine.

The purpose of the thesis is to study the behavior of time series and their construction on the basis of various autoregressive models, as well as the development of a software product for obtaining practical results and the choice of the best model for visualizing the data of short-term forecasting of demographic processes of population of Ukraine.

The object of the study is the statistical data on the processes of fertility, mortality and the general population of Ukraine converted into time series.

# ЗМІСТ

# ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ І ПОЗНАЧЕНЬ

AR(АР) – авторегресійне рівняння

ARMA(АРКС) – авторегресійне рівняння з ковзним середнім

ARIMA(АРІКС) – авторегресійне рівняння з інтегрованим ковзним середнім

SES(ЕЗ) – експоненціальне згладжування

DW – критерій Дарбіна-Уотсона

ACF(АКФ) – автокореляційна функція

PACF(ЧАКФ) – часткова автокореляційна функція

ME(КС) – ковзне середнє

SSE(СКП) – сума квадратів похибок моделі

СП – середня похибка прогнозу

СПП – середня похибка в процентах

СаПП – середня абсолютна похибка у процентах

$\boldsymbol{R^2}$ – коефіцієнт детермінації

AIC – інформаційний критерій Акайке

Theile – коефіцієнт Тейла

# ВСТУП

Сьогодні болючою постає глобальна проблема народонаселення в світі, тому керівництво провідних держав повинне з особливою увагою поставитись до цих питань. Адже нерегульовані демографічні процеси можуть призвести до перенаселення людей на одній території, звідки утвориться проблема видобутку необхідних продуктів життєдіяльності та природних ресурсів в цій місцевості. Основою суспільно-економічного розвитку для будь-якої країни є її народ. Відповідно місце держави в світовому ланцюгу існування залежить від рівня розвитку її населення. Необхідно уміло керувати геополітикою та сферами медицини, освіти, науки, суспільного дозвілля для того, щоб конвертувати здорову та освічену націю в соціально-економічну силу держави.

Але описані проблеми можна передбачити за допомогою дослідження і аналізу процесів народонаселення. Застосовуючи стандартні методи моделювання та прогнозування динаміки часових рядів, можна робити висновки щодо статево-вікової структури населення, кількості народжених немовлят на одну жінку, кількість розлучень та утворення подружь, смертність осіб по віковим категоріям, провести аналіз які чинники впливають або ж які етапи історії вплинули на ті чи інші зміни у формуванні структури народонаселення. Будуючи різноманітні моделі, за критеріями якості та адекватності моделей можна обрати модель, яка найкраще характеризує ситуацію поведінки певного статистичного ряду, а також зробити прогноз з різним рівнем тривалості.

# РОЗДІЛ 1
# ОСОБЛИВОСТІ ПРОБЛЕМАТИКИ СУЧАСНИХ ДЕМОГРАФІЧНИХ ПРОЦЕСІВ ТА ПОБУДОВИ ЇХ МАТЕМАТИЧНИХ МОДЕЛЕЙ

## 1.1 Аналіз проблематики сучасних демографічних процесів

Протягом всього періоду розвитку сучасного людства з’являлися різні проблеми, які ставали глобальними, адже впливали на подальше його існування та всього живого на планеті. Зокрема, це проблеми екології, енергетики, продовольства, проблеми війни і миру та інші, що набувають хронічного характеру[1]. Проте всі ці проблеми певною мірою є похідними глобальної проблеми демографічних змін суспільства. Безперечно, проблема народонаселення заважає соціально-економічному розвитку людського суспільства та будь-якій країні окремо, що й підкреслює необхідність її термінового дослідження та вирішення в найближчому майбутньому[2].

Демографія – це наука, що вивчає процеси становлення та зміни людського населення, керуючись емпіричними показниками: кількості населення, статево-вікової структури, приросту населення, динамічних рухів (міграція), народжуваності та смертності, тривалості життя, здійснення шлюбів та розлучень. Серед найінформативніших характеристик можна виділити декілька основних, якими ми і будемо користуватись в подальшому аналізі демографії в Україні: народжуваність, смертність та загальна чисельність населення.

Демографічна проблема почала прогресувати у 70-80-ті роки XX століття і має два протилежні аспекти. Одним з аспектів є явище демографічної кризи або ж депопуляція населення, що спостерігається в економічно розвинених країнах світу, та призводить до порушень відтворення населення та скорочення кількості жителів. Іншим аспектом поширеним серед держав, що розвиваються, є демографічний вибух. Дане

явище викликає ще більшу тривогу та характеризується швидкими темпами зростання кількості населення. Небезпека проявляється в тому, що відстала економіка і нерозвинена соціальна сфера не в змозі обернути цей ріст на благо свого розвитку. Також не менш гострими аспектами є неконтрольована урбанізація, криза великих міст світу, стихійна внутрішня та зовнішня міграція. Усі ці аспекти призводять до нерівномірного перерозподілу населення між різними регіонами світу.

Загострення питання урегулювання народонаселення збентежило більшість світових організацій по боротьбі з глобальними проблемами людства, а також було визнано усіма державами в наші дні.

### 1.1.1 Проблеми демографічних процесів у світі

Для того, щоб краще зрозуміти основні проблеми демографічних процесів в Україні, необхідно з’ясувати та проаналізувати першоджерела їхнього прояву у світі.

200 років тому на Землі проживало менше одного мільярда людей, сьогодні за підрахунками ООН, населення Землі складає 7 мільярдів. Останні оцінки свідчать, що сьогодні чисельність населення приблизно дорівнює 6,9% від загальної кількості людей, що коли-небудь народилися. Цей факт являється найбільш помітним росту населення: протягом тисячоліть темпи росту населення були незначні, проте останнього століття вони різко зросли. У період з 1900 по 2000 рік зростання світового населення було втричі більше, ніж за всю історію людства - збільшення від 1,5 до 6,1 мільярда всього за 100 років [3]. Динаміку росту кількості населення світу по регіонах за останні 200 років можемо спостерігати на Рисунку 1.1. Проте, якщо ми зосередимося на останніх десятиліттях, ми бачимо, що ця модель швидкого росту кількості більше не дотримується, оскільки річний темп приросту

населення останнім часом знижувався. У 1962 р. темпи зростання досягли максимуму на рівні 2,1%, і з того часу він знизився майже до половини, можна зробити висновок, що закінчився довгий історичний період прискореного зростання.

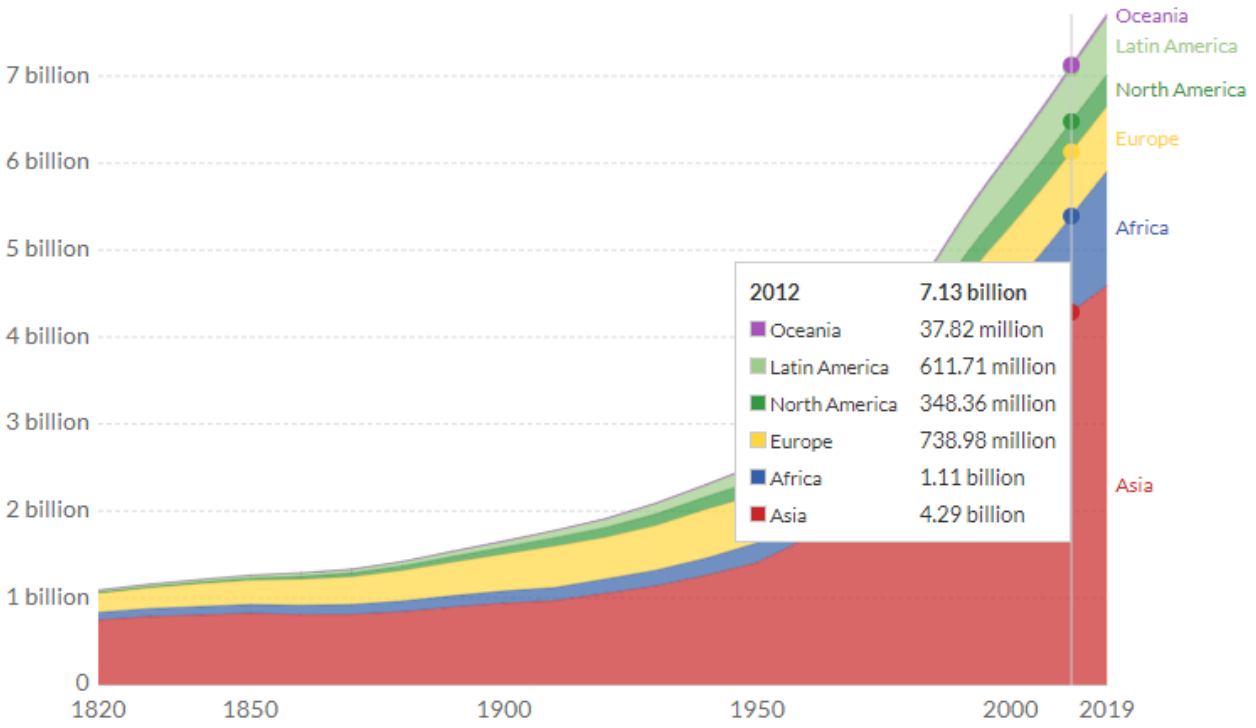



Рисунок 1.1 – Розподілення населення світу по регіонах

Проаналізувавши дані спостереження, світову історію становлення людства можна умовно розділити на три періоди, що характеризуються різними тенденціями зростання народонаселення. Першим є період, що тривав до початку сучасних часів, росту економіки, і був дуже тривалим віком повільного зростання населення. Другий період характерний зростанням рівня життя, медицини, економіки, освіти, він мав зростаючий темп збільшення кількості населення, який і продовжував зростати до 1962 року. Сьогодні другий період закінчився і почав розгортатися третій період, в якому темпи зростання населення падають.

Модель, яка враховує як змінювалось здоров’я людства за довгий час, динаміка зміни показників народжуваності та смертності населення,

соціально-економічні зміни в ході модернізації зниження дитячої смертності, структурні зміни в економіці і підвищення статусу й ролі жінки, повністю описує «демографічний перехід». Демографічний перехід, який ми спостерігаємо у всіх країнах по всьому світу, можна охарактеризувати п’ятьма етапами:

1. Перший етап є більшою частиною нашої історії. Упродовж тривалого часу, до сучасного зростання населення, рівень народжуваності був високий, але оскільки рівень смертності був також високий, то ми не спостерігали росту населення. Це описує реальність народонаселення усього світу протягом тисячоліть.
2. На другому етапі здоров’я поступово покращується, та рівень смертності зменшується. Оскільки рівень здоров’я населення вже зріс і народжуваність все ще залишається такою ж високою, як і раніше, то даний етап характеризувався стрімким ростом кількості населення. Історично це час, коли велика багатодітна сім’я являється звичайним явищем.
3. Третій етап характерний зниженням народжуваності в результаті соціальних змін: батьки розуміють, що прогрес в смертності дітей не такий високий, як було раніше, і тому вони вважають, що краще якісно виховати й забезпечити 1-2 дитини. У результаті відбуваються структурні зміни в економіці, які роблять дітей менш «економічно цінними», а жінки наділені соціальними і партнерськими зв’язками. Коефіцієнт народжуваності різко зростає.
4. Четвертий етап триває, коли приріст населення прямує до нуля, а рівень народжуваності співпадає з рівнем смертності.
5. П’ятий етап є етапом високого рівня розвитку. У даному етапі народжуваність зростає, але не на високому рівні. Внаслідок цього природний приріст населення складатиме нуль відсотків або, можливо, трішки вище.

Демографічний перехід описує зміни в ході соціально-економічної модернізації. Сьогодні країни повинні прагнути досягнути останнього етапу демографічного переходу, щоб вирішити глобальну проблему суспільства.

### 1.1.2 Проблеми демографічних процесів в Україні

Україна є п’ятою країною за кількістю населення в Європі після Росії, Англії, Франції, Італії та другою серед країн СНД і Прибалтики. За період отримання України незалежності і до сьогодні, у результаті соціально-економічних перетворень держави, необхідно визнати, що ситуація демографії держави значно погіршилась. Результатом цього є кризовий характер утворення значної депопуляції з погіршенням якісних характеристик здоров’я у поєднанні з динамічним рухом населення. За деякими даними чисельність населення країни у роки незалежності зменшилась на 5300 тисяч осіб, де депопуляція складає щонайменше 4600 тисяч.

Криза 90-х років, яка характерна соціально-економічними проблемами держави, прискорила тенденції зміни структури основних демографічних процесів. Адже до 2000-х років в країні неухильно і прискореними темпами знижувався рівень народжуваності та поширювалось явище бездітності. У 2005 році в Україні показник народжуваності становив 1,2 дитини на жінку та був одним за найгіршим серед європейців. Причиною таких даних є те, що держава  належним чином не забезпечувала сім’ї необхідними умовами, тобто був низький рівень економіки. У цілому сьогодні в українських сім’ях збережено традиційний спосіб відтворення та життя населення, де родина продовжує виконувати функції фізичного та морального забезпечення поколінь, проте є ці способи стають різноманітнішими і трансформуються як і в інших країнах Європи. Вони перетворюються в соціально-демографічні

проблеми та проявляються в утворенні неповних сімей, високій частці розлучень та малому коефіцієнта утворення законних шлюбів, соціально незабезпечених дітей-сиріт.

Фундаментальною особливістю становлення сучасної демографічної ситуації в Україні є низький рівень сфери забезпечення здоров'я та життєздатності суспільства. Високий рівень захворюваності людей працездатного віку та дітей, перевищення в значній мірі частки чоловічої смертності над жіночою більше, ніж в три рази, та великий показник смертності немовлят, демонструє недостатню організацію медицини в країні.

У 60-ті роки тривалість життя в Україні перевищувала тогочасний рівень Японії, Франції та Німеччини, на відміну від 2008 року, коли середня очікувана тривалість життя складає для чоловіків – 62,3 року, а для жінок – 73,8 року. Різниця між середньою тривалістю життя в Західній Європі та Україні у 2009 році становила 12,7 для чоловіків та 7,6 для жінок. Наслідком високого рівня чоловічої смертності є нездоровий спосіб життя, шкідливі і тяжкі умови праці та забруднення екології.

У 1992-2008 рр. досить високими темпами відбувалося зниження рівня народжуваності в Україні, що призводило до різних наслідків. Найбільш помітними вони було у Запорізькій, Донецькій, Дніпропетровській, Харківській, Луганській, Полтавській областях. У цих регіонах й зараз спостерігаються найнижчі показники. У результаті чого знизилась кількість дитячого населення майже на 1550 тисяч осіб.

Разом ці два фактори: зменшення народжуваності та збільшення смертності, призводять до подальшого росту від'ємного показника приросту населення. Так, у 2002–2003 рр. він становив –7,5 %, 2004 р. –7,0 %, за 2005 р. він зріс до – 7,6 %, а в 2006 р. зменшився до –6,4 %, і 2007 р. та 2008 р. складає 6,5 та 6,0 %, відповідно.

Також важливе значення в утворенні народонаселення в Україні займає значне зменшення кількості зареєстрованих шлюбів. Так за останні 10 років рівень шлюбності скоротився з 9,3 до 5,4 на 1 тисячу населення, а рівень

кількості розлучень в той самий період збільшився з 3,6 до 4,1 на 1 тисячу осіб[4]. Дана складова демографічних змін негативно впливає на відтворюваність українців.

Демографічна проблема України не являється одиничним випадком або ж суто нашим недоліком. Вона властива більшості розвинутим країнам, які подолали основні етапи демографічного переходу, проте не можуть подолати останній етап, де за рахунок розумного керування державними процесами регулювати популяцію свого суспільства у правильному ключі.

## 1.2 Математичні моделі та методи прогнозування демографічних процесів

Формулювання демографічного прогнозування зводиться до наукового дослідження основних показників демографічних процесів, таких як – кількість населення, статево-вікова структура, природній приріст населення, динамічні рухи (міграція внутрішня та зовнішня), народжуваність та смертність, тривалість життя, здійснення шлюбів та розлучень. Вони є основними характеристиками демографічних процесів і мають бути повністю обґрунтованими.

Результати прогнозування мають вагомий вплив на планування соціально-економічних процесів в державі, неурядових організаціях, компаніях та інших установах. Адже допомагають вирішити питання геополітики, перспектив виробництва, споживання товарів та послуг, планування житлового будівництва, розвитку соціальної інфраструктури, охорони здоров’я та освіти, пенсійних реформ та інші.

Демографічне прогнозування полягає у поєднанні деяких елементів моделювання та самого прогнозу. Дана методика називається потенційним аналізом, який у свою чергу поділяється на два основні типи: регіонально-

типологічний та соціально-інтегративний. Регіонально-типологічний тип методології розглядає потребу подолання жорсткої прив'язки демографічний процесів до територіальної диференціації. Областю вивчення якої є основні прогнозні характеристики в межах адміністративно-територіального устрою. Соціально-інтегративний тип базується на взаємодії аналізу та прогнозу населення, частки трудового потенціалу і процесами, які пов'язані з соціальною-економічною системою. У свою чергу, вони мають розглядатися як наслідки дії демографічних процесів[5].

Класифікацію методів та підходів демографічного прогнозування здебільшого проводять саме за ціллю прогнозування та проблемами, які необхідно вирішити:

- за регіональним охопленням (глобальні, регіональні, національні, територіальні);
- за часом (короткострокові, середньострокові, довгострокові);
- за призначенням (практичний, прогноз-застереження, нормативний, аналітичний, реалістичний, функціональний);
- за методикою розрахунку (екстраполяційний, кореляційно-регресійні моделі, метод марківських ланцюгів, тощо);
- за об'єктом (окремі вікові групи, народжуваність, смертність, гендерні групи);
- за ступенем деталізації (деталізовані та недеталізовані).

Завданням прогнозів-застережень є створення такої програми, яка формує інформацію про населення (смертність, народжуваність, міграційні процеси), для того, щоб суспільство могло своєчасно регулювати процеси, що несуть негативні наслідки.

У нормативних (або концептуальних) прогнозах спочатку обирається необхідне значення явища, яке необхідно спрогнозувати, з'ясовуються різні форми залежності у даному демографічному процесі, а потім розраховуються параметри незалежних змінних, які потрібні для досягнення необхідного результату. Концептуальні прогнози застосовуються для розробки стратегії

реалізації соціально-економічних заходів: підвищення рівня зайнятості населення, рівня освіта та здоров'я, популяризація бажаної кількості дітей на одну сім'ю. Потреби регулювання держави рівня народжуваності, смертності та міграційних процесів для контролю достатньої кількості природних ресурсів, необхідних для життєдіяльності людства.

Аналітичні прогнози є результатами наукових досліджень тенденцій закономірностей руху і репродукцію населення в перспективі. Їх, зазвичай, демонструють у вигляді оцінок параметрів майбутньої демографічної ситуації, яка створюється на основі припущення про незмінність утвореної тенденції. Оскільки аналітичні прогнози, як правило, є довгостроковими, вони є малоймовірними. Адже вказують лише ймовірні межі перспективних змін різних кількісних показників.

Реалістичні прогнози повинні точно демонструвати зміни у майбутньому розвитку народонаселення. Їх зазвичай використовують для розрахунку балансів робочої сили, народжуваності та смертності, переходу в інші вікові групи, а також статусу працездатної в пенсіонера. Для достовірності даного виду прогнозу в розрахунку застосовують всі параметри на довірчих інтервалах, які мають мінімальне та максимальне значення, проте значення показників буде коливатись у цих межах.

Функціональні прогнози використовують для прийняття рішень в соціальній, економічній, політичній та інших сферах діяльності державних та соціальних установ, і являють собою сукупність прогнозованої інформації про населення[5].

Демографічне прогнозування за методикою розрахунку класифікується на методи:

- екстраполяційний;
- аналітичний;
- марківських ланцюгів;
- регресійні моделі;
- когортно-компонентний;

У екстраполяційному методі використовуються дані про середньорічні абсолютні зміни чисельності населення за період або про середньорічні темпи приросту. Даний метод ґрунтується на прямому використанні лінійної та експоненціальної функцій. Прогнозований процес представляється як функція часу, в якому акумульовано дію інших факторів, що визначають його напрямок та інтенсивність. Якщо ці фактори вважати незмінними на весь період прогнозу, то можна розрахувати чисельність населення на будь-який період часу. За рахунок того, що розуміння цього методу просте він є досить популярний серед інших методів прогнозування, адже при виявленні наявного тренду можна спрогнозувати якими будуть дані у майбутньому. Проте й значним недоліком екстраполяційного методу є те, що він не враховує особливості розвитку статево-вікових груп населення. Тому для досліджень такого характеру, де на ці фактори необхідно зважати, цей метод небажано використовувати.

Аналітичний метод прогнозу складається на основі сформованого тренду за попередніми даними та функцією, яка найкраще та найдостовірніше описує даний тренд, будь-яка з цих функцій має емпіричний характер. Проте аналітичний метод так як і екстраполяційний застосовується здебільшого для короткострокових прогнозів та має однакові обмеження.

Метод марківських ланцюгів надає можливість враховувати в дослідженні прогнозування зміни вікових категорій людей, а також різні способи зникнення суб’єктів із області спостереження (смерть або внутрішні міграції). Формуються імовірності переходу одиниць та утворюються коефіцієнти, робиться припущення, що дані коефіцієнти не змінюються на якомусь інтервалі часу. Після чого формується така структура, яка залежить лише від матриці переходу, а не від початкових умов. Проте в цій структурі імовірнісна інтерпретація подій не є основною, важливою є соціально-економічна обумовленість та аналіз демографічних явищ у контексті й взаємодії із зміною якісного складу по кожному регіону.

Метод прогнозування за допомогою регресійних моделей краще за все використовувати у регіональному прогнозуванні. Оскільки даний метод в якості незалежної змінної використовує не час, а чисельно визначену матеріальну характеристику, що і є фактором. Оцінка здійснюється на основі вже відомих змін показників, що розглядаються, як факторні ознаки, та впливають на прогноз демографічних явищ. Формально цей метод прогнозування формується на основі побудови багатовимірних регресійних моделей, які є результатами аналізу множин кореляції та регресії. Важливим моментом у побудові прогнозу за допомогою регресійної моделі є не створення математичного апарату, а встановлення якісних взаємозв'язків між факторами, що формують їх інтенсивність[5].

Когортно-компонентний метод, який ще в літературі називають метод пересування вікових когорт, найчастіше використовують у прогнозуванні чисельності населення. Для реалізації даного метода необхідно мати дані про початкову чисельність та статево-вікову структуру. У процесі самого методу утворюються когорти за віком, і в майбутньому чисельність населення буде змінюватись за рахунок смертності та еміграції (тобто зменшуватись), народжуваності та імміграції (тобто збільшуватись).Усі ці процеси відбуваються тільки в певні моменти часу через однакові періоди, які називаються кроком прогнозування. Можливість отримати розподіл за статево-віковим чинником, а не тільки загальну кількість населення, і створює його перевагу порівняно з екстраполяційним та аналітичним методами[6].

Останнім часом у дослідників активно зростає увага до імовірнісного прогнозування. Можна виділити три основні альтернативні підходи, які набули практичного застосування: аналіз похибок прогнозу ex-post, аналіз динамічних рядів та експертна думка.

Аналіз похибок прогнозу ex-post є одним з альтернативних підходів імовірнісного прогнозування є результатом навчання на похибках історичних прогнозів, що були зроблені в минулому. За допомогою цих даних можна

визначати стійкі розподіли похибок та перетворювати результати історичного аналізу в довірчі інтервали, які потім будуть використані в новому прогнозі. Ананалізуючи прогнози зроблені в минулому, встановлюється величина і знак відхилення від тренда.

## 1.3 Огляд деяких комп'ютерних систем для побудови моделей прогнозування демографічних процесів

Метою багатьох провідних установ, що працюють над вирішенням глобальних проблем світу, було забезпечити якісною аналітичною інформацією пов'язаною з процесами народонаселення, соціально-економічними питаннями та коректним регулюванням демографічної політики кожну з держав. Тому в США в 1986 році було розроблено систему моделювання прогнозу «Spectrum» згідно з проектом «Полісі».

Україна була однією з перших східноєвропейських країн, яка випробувала та запровадила дану розробку в держустановах. Система моделювання «Spectrum» не лише дає результати прогнозування чисельності та регіонального складу населення, а й дає можливість передбачити наслідки певних змін, які є спричинені соціально-економічною державною політикою.

Комп'ютерна система прогнозування містить чотири основні модулі, що виконують різні типи прогнозів[7]:

1) «DemProj»(Demography) – програма, яка використовується здебільшого для демографічних процесів, а саме для прогнозування кількості населення, народжуваності, смертності в цілому та дитячої смертності, динамічних рухів(міграція);
2) «FamPlan»(Family Planning) – модуль, який застосовується для прогнозування процесу планування сім'ї, тобто сумарно рівня

народжуваності, завбачення небажаної вагітності та рівень забезпечення різновидів контрацепції в країні або регіоні.

3) «Rapid»(Resources for the Awareness of Population Impacts on Development) – програма прогнозування та аналізу статево-вікової групи, працездатного та пенсійного населення, що допомагатиме роботі по покращенню охорони здоров'я та освіти та інші;
4) «AIM»( AIDS Impact Model) – модуль для прогнозування та аналізу наслідків епідемії ВІЛ та СНІДу, у тому числі кількості людей, які живуть із хворобою, нових інфікованих, смертей за статтю та віком.

У основу комп'ютерної системи «Spectrum» покладено ідею імітаційного моделювання, що є одним з методів статистико-математичного моделювання та прогнозування. Воно полягає у з'ясуванні зміни якогось конкретного показника за допомогою факторів, значення яких змінюється вручну залежно від поставленої мети. Наприклад, необхідно дослідити зміни у чисельності населення після значних змін в статево-віковому складі та за рахунок збільшення сальдо динамічних рухів населення.

За допомогою кафедри статистики КНЕУ та Держкомстату України було створено додаткову нову версію одного з комп'ютерних модулів системи, яка повинна враховувати особливості сучасних демографічних процесів у нашій державі. Тому доданий модуль було адаптовано до даних про смертність населення України.

## Висновки до розділу 1 та постановка задачі дослідження

У даному розділі було виконану загальну характеристику демографічних процесів у світі та зроблено аналіз демографічних проблем в Україні.

Визначено основні напрямки розвитку держави щодо питань демографії у сферах економіки, медицини та інших галузях. Було проведено класифікацію прогнозів та розглянуто основні методи моделювання та прогнозування, які характерні для демографічних процесів. Також було описано комп'ютерні системи прогнозування, які використовуються.

Постановка задачі

1. Проаналізувати методи та моделі демографічних процесів і здійснити загальний їх опис.
2. Знайти необхідні статистичні дані за даною темою, які будуть необхідні для побудови описаних моделей та виконання прогнозування демографічних процесів.
3. Здійснити загальний опис основних математичних моделей прогнозування та вказати на їхні особливості.
4. Побудувати математичні моделі прогнозування різних видів та вказати їхні особливості під час експерименту.
5. Здійснити оцінку їх якості прогнозування демографічних процесів на основі відомих критеріїв оцінки прогнозування.
6. Продемонструвати порівняльний аналіз обраних моделей за вказаними критеріями оцінки, врахувавши загальну тенденцію особливостей демографії.
7. Зробити висновки даного експерименту та виробити поради стосовно подальших дій в демографічних процесах держави та напрямку її розвитку в сфері народонаселення.

# РОЗДІЛ 2
# ПОБУДОВА МАТЕМАТИЧНИХ МОДЕЛЕЙ ТА ПРОЦЕС ПРОГНОЗУВАННЯ

## 2.1 Формування структури моделі

Перш ніж розпочати роботу з математичними методами прогнозування та побудовою їх моделей необхідно якісно сформувати структуру моделі. Адже під час алгебраїчних та статистичних перетворень дана модель може поводити себе некоректно та давати результати, які не відповідають дійсності. На даному етапі необхідно скористатися всіма наданими джерелами інформації про процеси із метою вивчення їх особливостей. Оскільки модель може містити збурення, що впливають на процес, визначення наявності затримок на якісному та кількісному рівні, якщо це можливо визначення порядку процесу, а також наявність нелінійності і її характер. Якщо дослідження відбувається економічних процесів, то необхідно виявити чи наявна сезонність та чи присутній тренд, тобто зробити аналіз на інтегрованість. Виявивши в експерименті суттєву зміну рівня коливань на інтервалах часових рядів, зробити висновок про гетероскедастичність, а також перевірити гіпотезу про можливість введення коінтегрованості змінних. У результаті проведеного аналізу процесу можна побудувати ймовірну структуру моделі, яка і буде використовуватись в наступних засобах прогнозування.

### 2.1.1 Аналіз нелінійності

Варто зауважити, що перевірка на нелінійність може виконуватись на основі різних критеріїв, проте необхідно враховувати про всі їхні особливості. Результати різноманітних варіацій функцій, що мають характер нелінійності, не завжди є вдалими.

У економічних процесах допускається використання дисперсійного методу, що допомагає визначити наявність нелінійності, за допомогою функції:

$$\Psi_{zu}(t_1, t_2) = E_{u(t_2)}[E_{z(t_1)}[z(t_1)|u(t_2)] - E_{z(t_1)}[z(t_1)]]^2 ,$$

розв'язання даної функції відбувається за допомогою складного інтегрального рівняння.

Вирішити проблему нелінійності прогнозування процесів можна також за допомогою статистичного загального тесту Фішера (f-тест):

$$\hat{F} = \frac{\frac{1}{k-2}\sum_{i=1}^{k}\sum_{j=1}^{n_j} n_i(\overline{y_i} - \hat{y}_{ij})^2}{\frac{1}{k-2}\sum_{i=1}^{k}\sum_{j=1}^{n_j}(y_{ij} - \overline{y_i})^2} \tag{2.1}$$

де $k$ – число груп даних;

$n_i$ – кількість вимірів у групі;

$\overline{y_i}$ – групове середнє;

$\hat{y}_{ij}$ – оцінка по прямій регресії;

$n$ – загальне число вимірів.

Перефразувавши даний статистичний набір (2.1), отримаємо наступне відношення:

$$\hat{F} = \frac{\text{Відхилення середніх значень від прямої регресії}}{\text{Відхилення значень } y(k) \text{ від групових середніх}}$$

Припущення про лінійність процесу вважається хибних, якщо статистичний набір $\hat{F}$ із ступенями свободами рівними $\nu_1 = k - 2$, $\nu_2 = n - k$ дорівнює або більший рівня значущості[8].

### 2.1.2 Наявність коінтегрованості

Проблема стаціонарності процесу вирішується за допомогою розширеного тесту Дікі-Фуллера. Особливістю даного тесту є те, що значення залежної змінної з великими значеннями лагу вводиться в рівняння регресії, якого вистачить, щоб в тестові не застосовувати автокореляційні залишки. Дане рівняння набуває вигляду:

$$\Delta y(k) = a_0 + by(k-1) + c_1\Delta y(k-1) + c_2\Delta y(k-2) + \cdots + \\ + c_n\Delta y(k-n) + \varepsilon(k).$$

Середнє значення і компонент, за допомогою якого описується тренд утворюють окремий вид моделі, що тестується, та утворює форму критерію значимості.

Рівняння, яке тестується:

$$\Delta y(k) = by(k-1) + c_1\Delta y(k-1) + c_2\Delta y(k-2) + \cdots + \\ + c_n\Delta y(k-n) + \varepsilon(k),$$

де середнє значення рівне нулю та гіпотеза формулюється наступним чином:

$$H_0: b = 0 - \text{ ряд нестаціонарний;}$$

$$H_1: b < 0 - \text{ ряд стаціонарний.}$$

Якщо статистика $\frac{b}{SE_b}$ має значення менше від нуля та менше за критичне значення, яке було отримано із таблиці Діккі-Фуллера, то нульова гіпотеза не розглядається. Критичні значення рівнів значимості відповідно дорівнюють $\alpha = 1$ – -2,85 і $\alpha = 5$ – -1,95.

Якщо нульову гіпотеза є задовільною, то ряд $\{y(k)\}$ – являє собою випадкове блукання без зсуву, в рівнянні це константи.

У загальному вигляді цей критерій можна подати у вигляді обчислення модифікованого критичного значення з урахуванням розміру вибірки, що рівний $N$, та обчислюється за формулою:

$$\tau_\infty + \frac{\tau_1}{N} + \frac{\tau_1}{N^2}$$

де $\tau_\infty = -2.57\ (\alpha = 1)$ чи $\tau_\infty = -1.94\ (\alpha = 5)$;

$\tau_1 = -1.96\ (\alpha = 1)$ чи $\tau_1 = -0.398\ (\alpha = 5)$;

$\tau_2 = -10.04\ (\alpha = 1)$ чи $\tau_2 = 0\ (\alpha = 5)$ (значення $\tau$ табульовані Маккінноном, 1991).

Перевірка рівняння $\Delta y(k) = a_0 + by(k-1) + \varepsilon(k)$ з врахуванням можливої автокореляції залишків (як це було показано вище) базується на використанні того ж статистичного критерію, що і для рівняння без середнього, і тієї ж формули критичних значень, але з урахуванням наступних значеннях $\tau$:

$\tau_\infty = -3.43\ (\alpha = 1)$ чи $\tau_\infty = -2.86\ (\alpha = 5)$;

$\tau_1 = -6.00\ (\alpha = 1)$ чи $\tau_1 = -2.74\ (\alpha = 5)$;

$\tau_2 = -29.25\ (\alpha = 1)$ чи $\tau_2 = -8.36\ (\alpha = 5)$.

При наявності середнього та тренду застосовується така ж процедура, що і вище, але при наступних значеннях $\tau$:

$\tau_\infty = -3.96\ (\alpha = 1)$ чи $\tau_\infty = -3.41\ (\alpha = 5)$;

$\tau_1 = -8.35\ (\alpha = 1)$ чи $\tau_1 = -4.04\ (\alpha = 5)$;

$\tau_2 = -47.44\ (\alpha = 1)$ чи $\tau_2 = -17.83\ (\alpha = 5)$.

Розглянемо розширений тест Дікі-Фуллера.

Необхідно побудувати рівняння регресії, для використання тесту Дікі-Фуллерах[9]:

$$\Delta y(k) = a_0 + a_1 k + by(k-1) + \sum_{i=1}^{p} c_i\, \Delta y(k-i) + \varepsilon(k), \qquad (2.2)$$

де $a_0, a_1, b, c_i$ – невідомі коефіцієнти регресії.

Вище вказане рівняння (2.1) можна застосувати для реалізації тесту Діккі-Фуллера за умови, що коефіцієнти $c_i = 0, i = 1,2,\dots,p$, в іншому випадку потрібно використовувати загальний тест Дікка-Фуллера. У сфері економіки краще використовувати загальний тест Дікка-Фуллера із кількістю значень $p$ , які мали затримку в часі, меншою, ніж десять відсотків від числа спостережень, де $p < 0.1N$, де $N$ – довжина(потужність) часового ряду. Варто зауважити, що важливим моментом, коли використовується загальний та звичайний тест Діккі-Фуллера, є правильне формування структури моделі, адже важливу роль відіграє наявність параметрів $a_0$ і $a_1 k$.

Можна сформулювати евристичне правило, яке допоможе прискорити процес розв’язку даної задачі, а саме з візуалізації графіка зробити аналіз про наявність тренду. Якщо ж тренд відсутній, то в отриману модель (2.2)

потрібно включити тільки вільний член $a_0$, що є перетином. В іншому випадку, коли аналіз графіку вказує на присутність тренду, то в модель (2.2) вводяться наступні параметри $a_0$ і $a_1 k$.

Можна сформулювати наступні гіпотези на основі моделі (2.2):

$$H_0: b = 0 \text{ – ряд нестаціонарний: } y(k)\}\sim I(int), \text{ int } > 0;$$
$$H_1: b < 0 \text{ – ряд стаціонарний: } \{y(k)\}\sim I(0) \text{ ,int } = 0.$$

$H_0$ не розглядається, якщо маємо наступну нерівність оцінки коефіцієнта $\hat{b} < 0$ та виконаємо тестування на наявність одиничного кореня за допомогою обчислення $\tau$ – статистика Маккіннона , яка за абсолютною величиною є більшою за величину критичного значення даної статистики при регульованому рівні значущості $\alpha$.

Це можна записати наступним чином:

$$|\tau| = \left|\frac{\hat{b}}{SE_{\hat{b}}}\right| \geq |\tau_{крит}|$$

з рівнем значущості $\alpha$, де $SE_{\hat{b}}$ є стандартною похибкою оцінки $\hat{b}$.

### 2.1.3 Наявність коінтегрованості

Інтегрованість ряду – це перевірка динаміку ряду, який досліджується, на наявність тренду. Тестування можна здійснити певною кількістю способів, які будуть наведені в порядку збільшення рівня складності реалізації.

а) Графічний метод являє собою візуальне виявлення тренду на основі продемонстрованого графіку. Побудова графіку виконується наступним чином: по осі абсцис відкладається час, а по осі ординат

значення ряду. За характером поведінки кривої можна зробити про присутність або відсутність тренду часового ряду.

б) Метод середніх використовується після поділу досліджуваного ряду на два рівних підряди, для кожного з яких визначається середня величина $\overline{Y_1}$ та $\overline{Y_2}$. У результаті якщо знайдені значення відрізняються більше ніж на 10%, то можна констатувати факт наявності тренду в динамічному ряді.

в) Метод Стюарта та Кокса в дечому схожий із попередньо розглянутим, проте значною відмінністю є поділ того ж таки ряду на 3 рівні за кількістю рівнів групи. Особливої уваги заслуговує порівняння рівнів 1-ї та 3-ї груп. Якщо часовий ряд не можна порівну розділити на 3 частини, тобто кількість рівнів не ділиться на 3, то необхідно вилучити або ж добавити рівень, якого не вистачає.

г) У методі Мура та Валліса наявність тренду підтверджується тільки в тому випадку, якщо динамічний ряд не містить взагалі або ж містить в допустимій кількості фази, тобто зміну знаку під час ідентифікації абсолютного показника перетворення способом прив'язки.

д) Метод наборів використовується, коли кожний рівень ряду належить одному з двох типів. Наприклад, першому типові належить частина, яка менше середнього значення або ж медіани, а другому типові більше цих значень. Потім встановлюється, в уже створених типів послідовностях, кількість наборів $R$. Вони є послідовностями одного ж типу ряду, що межують з рівнями послідовності іншого типу. Якщо ряд не має загальну тенденцію росту або зниження рівнів, то кількість наборів є випадковою величиною з нормальним законом розподілу, за умови, що $n > 30$, або ж, якщо $n < 30$, то використовуємо розподіл Стьюдента. У результаті, якщо в змінах рівнів немає закономірностей, то випадкова величина належить довірчому інтервалу:

$$(\bar{R} - t\sigma) \leq R \leq (\bar{R} + t\sigma) \qquad (2.3)$$

де $t$ – довірчий коефіцієнт для допустимого рівня ймовірності в нормальному розподілі, а в розподілі Стьюдента із ступенем свободи рівним $k = (n-1)$;

$\bar{R}$ – середня кількість наборів в ряді, що знаходиться за формулою:

$$\bar{R} = \frac{n+1}{2} \tag{2.4}$$

$\sigma$ – середнє квадратичне відхилення кількості наборів в ряді, що розраховується за формулою:

$$\sigma = \frac{\sqrt{n-1}}{2} \tag{2.5}$$

Підставимо виведення показників (2.4) та (2.5) в довірчий інтервал (2.3) та отримаємо наступне перетворення:

$$\frac{(n+1-t\sqrt{n-1})}{2} \leq R \leq \frac{(n+1+t\sqrt{n-1})}{2}$$

Отже, якщо встановлена кількість наборів ряду не входить в довірчий інтервал, то у динамічному ряді наявний тренд, якщо значення $R$ належить довірчому інтервалу, то можна констатувати відсутність тренду.

### 2.1.4 Аналіз на гетероскедастичність

Існує декілька тестів за допомогою яких можна дослідити часовий ряд на наявність гетероскедастичності.

Розглянемо Тест Бройша-Пагана (Годфрі).

Маємо наступну лінійну регресію:

$$y(k) = X^T(k)\beta + \varepsilon(k), \tag{2.6}$$

де $X^T(k) = [1\ x_2(k)\ x_3(k)\ \dots\ x_r(k)]$.

Нехай гетероскедастичність приймає наступний вигляд:

$$E[\varepsilon(k)] = 0, \qquad \forall k,$$

$$var[\varepsilon(k)] = E[\varepsilon^2(k)] = \sigma_\varepsilon^2 = h\big(\alpha z^T(k)\big), \tag{2.7}$$

де $z^T(k) = [1\ z_2(k)\ z_3(k)\ \dots\ z_p(k)]$ – вектор відомих змінних;

$\alpha = [\alpha_1\ \alpha_2\ \ \alpha_3\ \ \dots\ \alpha_p\ ]$ – вектор невідомих коефіцієнтів;

$h(\cdot)$ – довільна невизначена функція, що приймає лише значення більші нуля.

Нульова гіпотеза гомоскедастичності:

$H_0 : \alpha_2 = \alpha_3 = \ \dots\ = \ \alpha_p = 0,$ що означає $\sigma_\varepsilon^2 = h(\alpha_1) = const.$

За цією нульовою гіпотезою можна оцінювати коефіцієнти виразу (2.6) з використанням методу найменших квадратів, припустивши, що розподіл збурень у правій частині рівняння є нормальним. Процедуру застосування тесту на гетероскедастичність можна легко зобразити у вигляді нескладного алгоритму:

1) Спочатку необхідно здійснити оцінку параметрів початкової моделі (2.6), використовуючи простий МНК, а також сформувати масив залишків $e(k) = y(k) - X^T(k)\beta$ та обчислити дисперсії $\sigma_\varepsilon^2 = \sigma_e^2 = N^{-1}\sum e^2(k)$.
2) Знайти оцінки регресії $e^2(k)/\sigma_e^2$ на $z(k)$ за допомогою ЗМНК та обчислити значення похибки ESS за формулою $ESS = \beta^T X^T X\beta - N\mu_y^2$, де $\mu_y^2$ – середнє значення послідовності $\{y(k)\}$.
3) Застосовуючи нуль-гіпотезу $H_0$: $\frac{1}{2}ESS \leftrightarrow \chi^2(p-1)$ . Отже, гіпотеза стосовно гомоскедастичності не береться до уваги, якщо $ESS/2$ перевищує вибране критичне значення із розподілу $\chi^2$.

4) Асимптотично еквівалентним підходом, який є легше реалізувати, є оцінювання регресії $e^2(k)$ на $z(k)$. Величина $NR^2$, обчислена для цієї регресії, буде мати в асимптотичний розділ $\chi^2(p-1)$ при розгляді нульової гіпотези.

Варто зауважити, що зв'язок між різними видами похибок регресійної моделі визначається за виразом: $\left(y^T y - N\mu_y^2\right) = \left(\beta^T X^T X\beta - N\mu_y^2\right) + e^T e$, або $TSS = ESS + RSS$,

де $TSS$ –загальна похибка регресії;

$RSS$ – сума квадратів похибок моделі.

Використовуючи продемонстрований тест необхідно знати змінні $z$, які створюють гетероскедастичність, але немає потреба завчасно знати функціональну структуру гетероскидастичність. У дослідженнях кандидати в змінні $z$ можуть бути вибрані з векторів регресорів $x(k)$. Якщо дане припущення підтверджується, то алгоритм застосування даного теста відповідає алгоритму тесту Уайта[8].

Розглянемо Тест Уайта.

Для тестування гетероскедастичності у цьому тесті потрібно сформувати додаткову модель регресії для квадратів залишків, що будуть генеруватись з використанням методу найменших квадратів. Дана модель регресія буде складатися з ненадлишкових регресорів на всій області регресорів, що містять квадрати, взаємні добутки регресорів та власне їх, а також модель міститиме константу в правій частині. Побудуємо регресію у вигляді:

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + \varepsilon(k),$$

де $[1\ x_1\ x_2\ ]^T$ - вектором незалежних змінних.

Множина всіх регресорів, їх квадратів та взаємних добутків, ненадлишкових змінних матиме наступну форму: $[1\ x_1\ x_2\ {x_1}^2\ {x_2}^2\ x_1\ x_2\ ]$.

Якщо модель має гетероскидастичні процеси, то $NR^2 \leftrightarrow \chi^2(q)$. Тобто $NR^2$ може мати асимптотичний розподіл з параметром $q$, який означає кількість регресорів без константи і в даному випадку буде дорівнювати 5, $\chi^2(5)$ . Даний тест може виявити гетероскедастичність часового ряду, проте не демонструє структуру початкової моделі й спосіб знаходження її параметрів. Але оцінку параметрів можна знайти за допомогою методу найменших квадратів.

Також значним недоліком даного тесту є те, що розподіл $\chi^2$ може мати велику кількість рівнів свободи, що сприятиме значному збільшенню параметру $q$ і відповідно зменшуватиме якість даного тесту. Іноді за рахунок вилучення взаємних добутків із загальної множини регресорів та шляхом вводу в регресію їх квадратів, зменшують значення параметра $q$ [8] .

Також існує тест Голдфельда-Квандта, який використовують у випадках, коли є одна змінна, зокрема з числа регресорів, що породжує гетероскедастичність. Якість даного тесту залежить від об'єму спостережень, що не розглядаються в дослідженні.

## 2.2 Математичні методи прогнозування

Розглянемо простий приклад рівняння АР(1):

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k),\ E[\varepsilon(k)] = 0 \qquad (2.8)$$

де $\{\varepsilon(k)\}$ – послідовність білого шуму з нульовим середнім.

Збільшимо незалежну змінну $k$, яка має зміст часу, на одиницю і запишемо рівняння знову:

$$y(k+1) = a_0 + a_1 y(k) + \varepsilon(k+1)$$

При відомих $a_0$, $a_1$ можемо знайти математичне сподівання $y(k+1)$ до дискретного моменту часу $k$.

$$E_k[y(k+1)] = a_0 + a_1 E_k[y(k)] = a_0 + a_1 y(k)$$

адже $y(k)$ в момент часу $k$ є відомою константою.

Аналогічно з виразом (2.8) запишемо для моменту $k\ +2$:

$$y(k+2) = a_0 + a_1 y(k+1) + \varepsilon(k+2)$$

і з допомогою нескладних арифметичних дій знайдемо математичне сподівання $y(k+2)$:

$$E_k[y(k+2)] = a_0 + a_1 E_k[y(k+1)] = a_0 + a_1 E_k[a_0 + a_1 y(k)] = a_0 + \\ + a_0 a_1 + a_1^2 y(k)$$

Отже, тепер ми можемо записати вираз для загального випадку прогнозування на $s$ кроків:

$$E_s[y(k+s)] = a_0\left(\sum_{i=o}^{s-1} a_1^i\right) + a_1^s y(k) = a_0 \sum_{i=o}^{s-1} a_1^i + a_1^s y(k) \qquad (2.9)$$

Вираз (2.9), що ми отримали, є функцією прогнозування на довільне число кроків. Прогнозування являє собою збіжний процес, якщо $|a_1| < 1$, отже:

$$lim_{s\to\infty} E_k[y(k+s)] = \frac{a_0}{1-a_1} \qquad (2.10)$$

де $a_1$– знаменник геометричної прогресії.

Із отриманого виразу (2.10), можна зробити висновок , для будь-якого стаціонарного процесу АР або ж АРКС, оцінка умовного прогнозу асимптотично (s → ∞) збігається до безумовного середнього.

Доцільно обчислити похибку прогнозу за умові, що умовне математичне сподівання $E[\varepsilon(k)] = 0$:

$$f_k(s) = y(k+s) - E_k[y(k+s)].$$

Розрахуємо похибку прогнозу, де $s = 1$:

$$f_k(1) = y(k+1) - E_k[y(k+1)] = a_0 + a_1 y(k) + \varepsilon(k+1) - a_0 - a_1 y(k) = \\ = \varepsilon(k+1).$$

Розрахуємо похибку прогнозу, де $s = 2$:

$$f_k(2) = y(k+2) - E_k[y(k+2)] = \\ = a_0 + a_1[a_0 + a_1 y(k) + \varepsilon(k+1)] + \varepsilon(k+2) - E_k[y(k+2)] = a_0 + \\ + a_0 a_1 + a_1^2 y(k) + a_1 \varepsilon(k+1) + \varepsilon(k+2) - a_0 - \\ - a_0 a_1 - a_1^2 y(k) = \varepsilon(k+2) + a_1 \varepsilon(k+1).$$

Отже, можна записати рівняння загального виду похибки прогнозу з довільною кількістю кроків прогнозування:

$$f_k(s) = \varepsilon(k+s) + a_1 \varepsilon(k+s-1) + a_1^2 \varepsilon(k+s-2) + \cdots + a_1^{s-1} \varepsilon(k+1).$$

Тепер можна розрахувати дисперсію похибки прогнозування, врахувавши математичне сподівання $E_k[f_k(s)] = 0$. Отже оцінка прогнозу,

яка обчислюється за виразом (2.9), є незміщеною. Маємо наступний вираз дисперсії похибки:

$$Var[f_k(s)] = \sigma^2[1 + a_1^2 + a_1^4 + a_1^6 + \cdots + a_1^{2(s-1)}],$$

тобто дисперсія є функцією $s$. Асимптотичне значення дисперсії похибки прогнозу для стаціонарного процесу:

$$\lim_{s\to\infty} Var[f_k(s)] = \frac{\sigma^2}{1 - a_1^2}$$

де $a_1^2$ – знаменник геометричної прогресії.

Ковзне середнє використовується в економіці та фінансах для згладжування цінових рядів, короткострокових коливань, акцентуючи увагу на основних тенденціях і циклах. При обчисленні ковзного середнього значення функції розраховується кожен раз заново, враховуючи при цьому кінцеву множину попередніх значень. Ковзне середнє певним чином «рухається» по часовому ряду. Оскільки при розрахунку береться середнє значення набору величин, то воно певним чином перебуває в постійному русі відносно часу. У процесах де закономірність динамічних рухів мало помітна, ковзне середнє змінюється у горизонтальному інтервалі.

Загальний вигляд формули для зваженого ковзного середнього:

$$MA(k) = \frac{\sum_{i=1}^{N} w_i \cdot y(k - i + 1)}{\sum_{i=1}^{N} w_i}$$

де $N$ – розмір вікна ковзного середнього;

$w_i$ – вагові коефіцієнти;

$y$ – часовий ряд вхідних даних.

Побудуємо функцію прогнозування моделі АРКС(2,1):

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \varepsilon(k) + \beta_1 \varepsilon(k-1),$$

де $\varepsilon(k)$ – білий шум з нульовим середнім;

$y(0) = y_0$– відома початкова умова.

Збільшимо незалежну змінну $k$, яка має зміст часу, на одиницю і запишемо рівняння знову:

$$y(k+1) = a_0 + a_1 y(k) + a_2 y(k-1) + \varepsilon(k+1) + \beta_1 \varepsilon(k)$$

Тепер можемо знайти математичне сподівання $y(k+1)$ до дискретного моменту часу $k$, рівняння якого матиме вигляд:

$$E_k[y(k+1)] = a_0 + a_1 y(k) + a_2 y(k-1) + \beta_1 \varepsilon(k),$$

де $\varepsilon(k)$ відома величина на момент часу *k*.

$$E_k[\varepsilon(k+j)] = 0, \forall j > 0$$

Запишемо рівняння для *k+2*:

$$y(k+2) = a_0 + a_1 y(k+1) + a_2 y(k) + \varepsilon(k+2) + \beta_1 \varepsilon(k+1)$$

Математичне сподівання для k+2:

$$E_k[y(k+2)] = a_0 + a_1 E_k[y(k+1)] + a_2 E_k[y(k)] =$$
$$= a_0 + a_1[a_0 + a_1 y(k) + a_2 y(k-1) + \beta_1 \varepsilon(k)] + a_2 y(k) =$$
$$= a_0 + a_0 a_1 + a_1^2 y(k) + a_1 a_2 y(k-1) + a_1 \beta_1 \varepsilon(k) + a_2 y(k) =$$
$$= a_0(1 + a_1) + (a_1^2 + a_2) y(k) + a_1 a_2 y(k-1) + a_1 \beta_1 \varepsilon(k)$$

Знайдемо математичне сподівання на три кроки вперед:

$$E_k[y(k+3)] = a_0 + a_1 E_k[y(k+2)] + a_2 E_k[y(k+1)] =$$
$$= a_0(1 + a_1 + a_1^2 + a_2) + (a_1^3 + 2a_1 a_2)y(k) +$$
$$+(a_1^2 a_2 + a_2^2)y(k-1) + \beta_1(a_1^2 + a_2)\varepsilon(k).$$

Отримавши попередні формули умовних математичних сподівань і провівши аналіз, можна вивести формулу для загального умовного математичного сподівання на $s$ кроків, яка матиме наступний вигляд:

$$E_k[y(k+s)] = a_0 + a_1 E_k[y(k+s-1)] + a_2 E_k[y(k+s-2)] \quad (2.11)$$

Якщо корені отриманого характеристичного рівняння (2.11) знаходяться всередині одиничного кола, то оцінка прогнозу асимптотично збігається до безумовного середнього значення

$$\lim_{s\to\infty} E_k[y(k+s)] = \frac{a_0}{1-a_1-a_2},$$

а для довільного процесу АРКС (p,q) оцінку умовного прогнозу можна записати як:

$$E_k[y(k+s)] = a_0 + \sum_{i=1}^{p} a_i E_k[y(k+s-i)]$$

Модель ARMAX( p,q,d) у загальному вигляді:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + mv(k) + \sum_{j=1}^{q} b_j \cdot mv(k-j) + \sum_{s=1}^{d} c_s x_s, \quad (2.12)$$

де $p$ – порядок авторегресійної частини;

$q$ – порядок ковзного середнього;

$d$ – кількість включених пояснюючих змінних.

ARMAX(AutoRegressive Moving Average model with eXogenous inputs model) модель, яка характерна наявністю екзогенного(зовнішнього) фактору.

У виразі (2.12) $\sum_{s=1}^{d} c_s x_s$ представляє собою лінійну комбінацію зовнішніх пояснюючих змінних $x_1, \dots, x_d$. У реальних демографічних процесах вони можуть являти собою додаткові чинники впливу на загальну картину демографії, такі як вплив ВВП країни, фінансування сфери медицини, покращення освіти, ріст кількості освічених людей в країні, роль жінки в суспільстві.

На основі аналізу сумісної кореляції вихідного сигналу $y$ та $x_s$, що записується наступним чином $correl(y, x_s) = r_{y,x_s}$, було вирішено включити в модель ARMAX відповідну пояснювальну змінну $x_s$. Якщо $r_{y,x_s} > 0.5$, то змінну $x_s$ необхідно включати.

У загальному випадку до складу ARMAX рівняння окрім регрессорів $x_1, \dots, x_d$ також можуть включатися лагові змінні $x_s(k-m)$. Для того, щоб включити відповідну авторегресійну частину регрессора $x_s$ необхідно виконати аналіз ЧКФ$(y, x_s)$[11].

## 2.3 Критерії вибору кращої моделі та прогнозу

Перш ніж формувати структуру математичної моделі прогнозування необхідно вибрати, яка з моделей буде краще описувати характер поведінки

динамічних процесів, адже процеси в різних галузях їхнього прояву мають різні особливості. Для вирішення даної проблеми існують різноманітні критерії адекватності моделі, які допомагають вірно обрати модель прогнозування та оцінити її актуальність щодо досліджуваного процесу. До відомих критеріїв належать: статистичні параметри ( t – статистика Стьюдента, коефіцієнт детермінації $R^2$, сума квадратів похибок моделі(SSE)), інформаційний критерій Акайке (AIC), статистика Дарбіна-Уотсона(DW), статистика Фішера, коефіцієнт Тейла (Theile).

Після того, як була вибрана вже модель і побудований прогноз необхідно визначити чи є цей прогноз доцільний. Отримавши прогноз різними математичними моделями, кожна модель має свої переваги та недоліки в інформативності зробленого прогнозування. Тому існують оцінки точності прогнозу, що і характеризують його: середньоквадратична похибка(СКП), середня похибка прогнозу(СП), середня похибка в процентах(СПП), середня абсолютна похибка у процентах(СаПП), максимальна абсолютна похибка(МАП), мінімальна абсолютна похибка (МіАП).

### 2.3.1 Критерії адекватності моделі

Критерії адекватності моделі дозволяють оцінити окремо значущість коефіцієнтів математичної моделі в статистичному сенсі, визначити інтегральну похибку моделі стосовно вихідного часового ряду, встановити наявність кореляції між значеннями похибки моделі, адже вони мають бути не корельованими, а також визначити ступінь адекватності моделі фізичному процесу в цілому. Розглянемо деякі з них.

Розглянемо коефіцієнт детермінації $R^2$.

Оскільки мірою інформативності часового ряду найчастіше використовують дисперсію, то $R^2$ базується саме на цій ідеї. Обчислюється коефіцієнт детермінації за формулою:

$$R^2 = \frac{var(\hat{y})}{var(y)} = 1 - \frac{SSE}{SST}$$

де $var(\hat{y})$ – дисперсія частини часового ряду основної змінної рівняння;

$var(y)$ – вибіркова дисперсія цієї змінної;

$SSE = \sum_{k=1}^{N}[y(k) - \hat{y}(k)]^2$ – сума квадратів похибок (залишків) моделі;

$SST = \sum_{k=1}^{N}[y(k) - \bar{y}]^2$ – загальна сума квадратів $\bar{y}$ – середнє значення.

Модель вважається кращою, якщо $R^2 \to 1$.

Розглянемо критерій, що являється сумою квадратів похибок моделі(SSE)

Обчислюється за формулою:

$$\text{SSE} = \sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N}[\hat{y}(k) - y(k)]^2 \to min$$

де $\hat{y}(k) = \hat{a}_0 + \hat{a}_1\hat{y}(k-1) + \hat{a}_2\hat{y}(k-2) + \hat{b}_1x(k) + b_2z(k)$;

$y(k)$ – вимірювання;

$N$ – довжина вибірки.

Отже, для отримання найкращої моделі з усіх можливих кандидатів потрібно обрати ту, в якої значення $\sum e^2(k)$ є найменшим.

Розглянемо критерій Акайке (AIC)

Даний критерій бере до уваги $\sum e^2(k)$, N-вимірів та число параметрів моделі. Обчислюється за наступною формулою:

$$AIC = Nln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n$$

де $n = p + q + 1$ – число параметрів моделі, які оцінюються за допомогою статистичних даних ($p$- число параметрів авто регресійної частини моделі; $q$ - число параметрів ковзного середнього; одиниця з'являється тоді, коли оцінюється зміщення (або перетин), тобто $a_0$).

Оскільки даний метод враховує значення $\sum e^2(k)$, то при виборі кращої моделі-кандидата, необхідно прагнути мінімального значення $AIC$. Інформативнішим його робить наявність у формулі таких значень як N-довжина вибірки та число оцінюваних параметрів.

Розглянемо такий критерій, як Статистика Дарбіна-Уотсона (DW).

Обчислюється за формулою:

$$DW = 2 - 2\rho,$$

де $\rho = \frac{E[e(k)e(k-1)]}{\sigma_e^2}$ – коефіцієнт кореляції між сусідніми значеннями похибки;

$\sigma_e^2$ – дисперсія послідовності похибок $\{e(k)\}$.

Таким чином, при повній відсутності кореляції між похибками $DW = 2$ – це ідеальне значення. Граничними значеннями для $DW$ є 0 (при $\rho = 1$) та +4 (при $\rho = -1$).

Розглянемо коефіцієнт Тейла (Theile)

Обчислюється за формулою:

$$U = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i)^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i)^2}}$$

Маємо допустимі значення коефіцієнта $0 \leq U \leq 1$. Модель не може використовуватись у прогнозуванні при $U = 1$, інакше, коли $U = 0$ це означає, що ряд прогнозу співпадає з реальними рядом, тобто в такому випадку модель найкраще описує реальний процес.

### 2.3.2 Оцінка точності прогнозу

Після побудови та виборі найкращої моделі, необхідно обчислити оцінку прогнозу ряду та виявити поведінку ряду.

1) Середньоквадратична похибка

$$\text{СКП} = \sqrt{\frac{1}{S}\left(y(k+s) - \hat{y}(k+s,k)\right)^2}$$

2) Середня похибка прогнозу:

$$\text{СП} = \frac{1}{S}\sum_{i=1}^{S} y(k+s) - \hat{y}(k+s,k)$$

3) Середня похибка в процентах:

$$\text{СПП} = \frac{1}{S}\sum_{i=1}^{S} \frac{y(k+s) - \hat{y}(k+s,k)}{y(k+s)} \times 100\%$$

4) Середня абсолютна похибка у процентах:

$$\text{АСПП} = \frac{1}{S}\sum_{i=1}^{S}\frac{|y(k+s) - \hat{y}(k+s,k)|}{|y(k+s)|} \times 100\%$$

5) Максимальна абсолютна похибка:

$$\text{МАП} = \max\{|y(k+1) - \hat{y}(k+1,k)|, \dots, |y(k+s) - \hat{y}(k+s,k)|\}.$$

6) Мінімальна абсолютна похибка:

$$\text{МіАП} = \min\{|y(k+1) - \hat{y}(k+1,k)|, \dots, |y(k+s) - \hat{y}(k+s,k)|\}$$

## Висновки до розділу 2

У даному розділі було розглянуто побудову математичної моделі та всі рівні перевірки особливостей моделі та її критерії адекватності.

Перш ніж будувати саму модель, необхідно якісно сформувати структуру моделі, а саме дослідити її на лінійність, стаціонарність, гетероскедастичність та інтегрованість, за допомогою відомих тестів, що й було продемонстровано.

Було розглянуто моделі авторегресійного рівняння та авторегресійного рівняння з ковзним середнім. Помітною перевагою ARMA є якість моделювання та прогнозування, порівняно з AR. Проте процес прогнозування ARMA займає значно більше часу та завантаженості даних, ніж AR. Перевагою AR є наявність лише одного типу порядку моделі, коли в ARMA їх два, що дає змогу автоматично обирати кращу модель, а в моделі з ковзним середнім необхідно робити це власноруч. Проте це не складає великої проблеми, адже за структурою модель є не складною і в ній досить легко розібратись.

Для оцінки адекватності математичної моделі та її прогнозу використовуються найпоширеніші критерії: коефіцієнт детермінації $R^2$, сума квадратів похибок моделі, коефіцієнт Тейла, середньоквадратична похибка.

# РОЗДІЛ 3
# ПОБУДОВА МАТЕМАТИЧНИХ МОДЕЛЕЙ ДЕМОГРАФІЧНИХ ПРОЦЕСІВ В УКРАЇНІ

## 3.1 Програмна реалізація та її архітектура

Під час написання дипломної роботи було розроблено та практично реалізовано програмний продукт для створення математичних моделей та короткострокового прогнозу на основі часових рядів, статистичними даними якої слугували основні показники демографічних процесів в Україні.

Основною ціллю створення даного програмного продукту було створення універсального функціоналу, з допомогою якого можна досліджувати поведінку динамічних рядів, будувати математичні моделі та прогнозувати різноманітні процеси. У даній роботі програма була застосована для аналізу демографічних процесів в Україні для розробки подальших та плануванні коректного регулювання народонаселення державою.

Продемонструємо алгоритм функціонування програмного продукту у вигляді схеми(Рисунок 3.1):

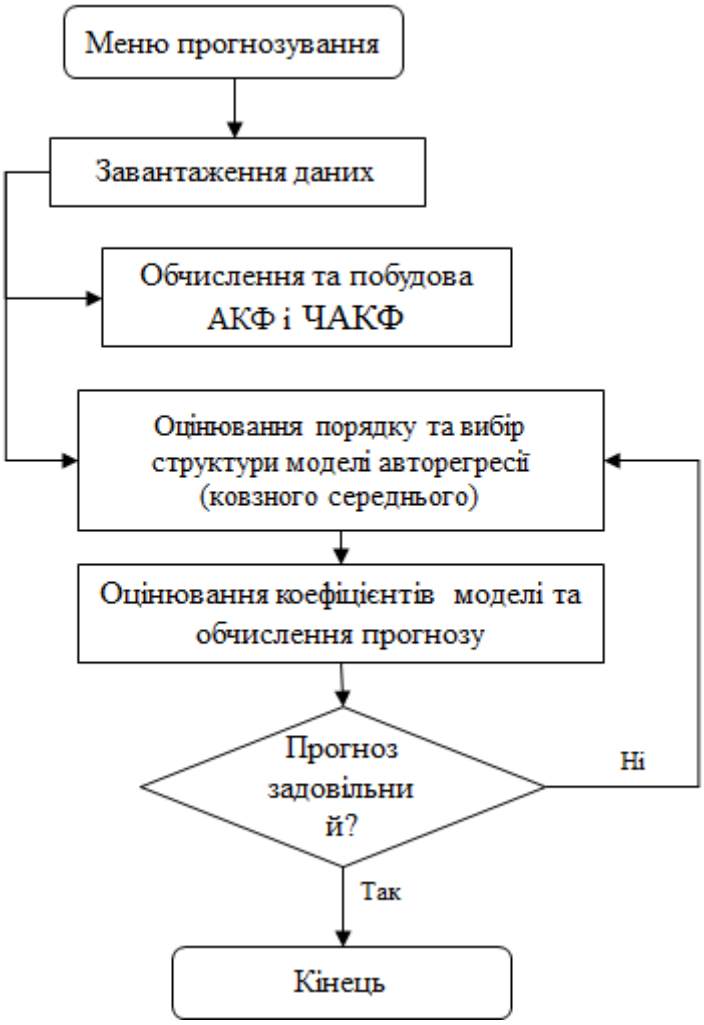



Рисунок 3.1 – Алгоритм функціонування програми

## 3.2 Аналіз вибору програмного середовища для реалізації та зручного функціонування

Під час виконання дипломної роботи для вибору програмного середовища було запропоновано декілька варіантів: Python, C#, Matlab, R. У результаті перевагу було віддано на користь мови програмування Python, а саме розробка проводилась в середовищі Jupyter Notebook Google Colaboratory з останньою версією Python 3.7.3.

Python – одна з найпопулярніших мов програмування високого рівня зі строгою динамічною типізацією. Перевагами даної платформи є її можливість використання в діалоговому режимі, що дуже зручно при розробці експериментів та досліджень, зручна для розв'язування математичних проблем та моделювання, а також дуже інформативна візуалізація даних різного типу, пристосування платформи розробки до роботи з матрицями та великими масивами даних без нагальної необхідності попереднього виділення пам'яті та задання розмірності і типу даних.

Однією з головних переваг серед інших середовищ є наявність зручної бібліотеки Statsmodels, в якій вже наявні стандартні підходи до математичного моделювання та прогнозування. Statsmodels є пакетом Python, який дозволяє користувачам досліджувати дані, оцінювати статистичні моделі і виконувати статистичні тести. Великий перелік описової статистики, статистичних тестів, графічних функцій і статистики результатів доступний для різних типів даних і кожного оцінювача. Він доповнює модуль статистики SciPy. Statsmodels є частиною наукового стеку Python, який орієнтований на аналіз даних, наукb даних і статистику. Statsmodels побудований поверх чисельних бібліотек NumPy і SciPy, інтегрується з Pandas для обробки даних і використовує Patsy для інтерфейсу R-подібної формули. Графічні функції засновані на бібліотеці Matplotlib. Statsmodels надає статистичний сервер для інших бібліотек Python. Statmodels є вільним програмним забезпеченням, що випускається під ліцензією модифікованої BSD (3-клаузи).

## 3.3 Побудова математичних моделей та короткострокового прогнозування на основі статистичних даних

### 3.3.1 Регресійні моделі та прогноз народжуваності України

Однією з основних характеристик демографічного стану держави є її народжуваність, тому в наступному підрозділі буде розглянуто особливості поведінки даного показника вибірка якого налічує 161 елемент помісячно на проміжку 2003-2016 років.

Зобразимо стан народжуваності за період 2003-2016 рр. на Рисунку 3.2:

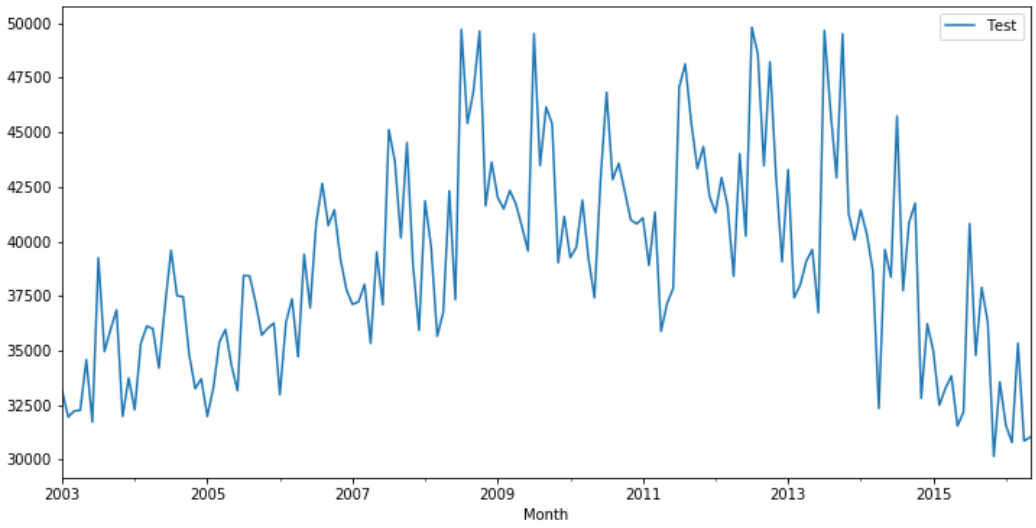



Рисунок 3.2 – Реальний стан народжуваності за період 2003-2016 рр.

Для кращого вибору порядку регресійних моделей побудуємо ACF(АКФ) - автокореляційну функцію та PACF(ЧАКФ) – часткову автокореляційну функцію на наступних рисунках відповідно Рисунок 3.3 та Рисунок 3.4.

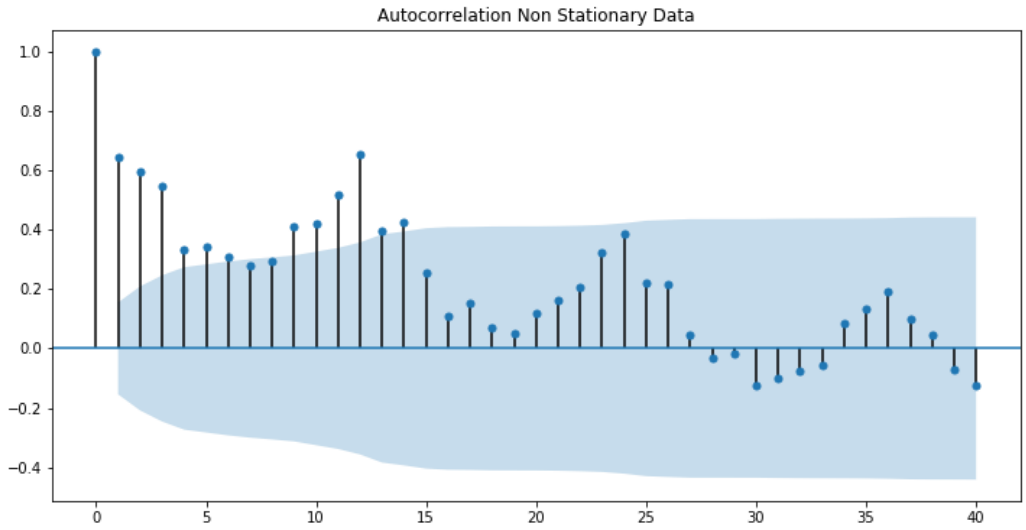



Рисунок 3.3 – Автокореляційна функція народжуваності України

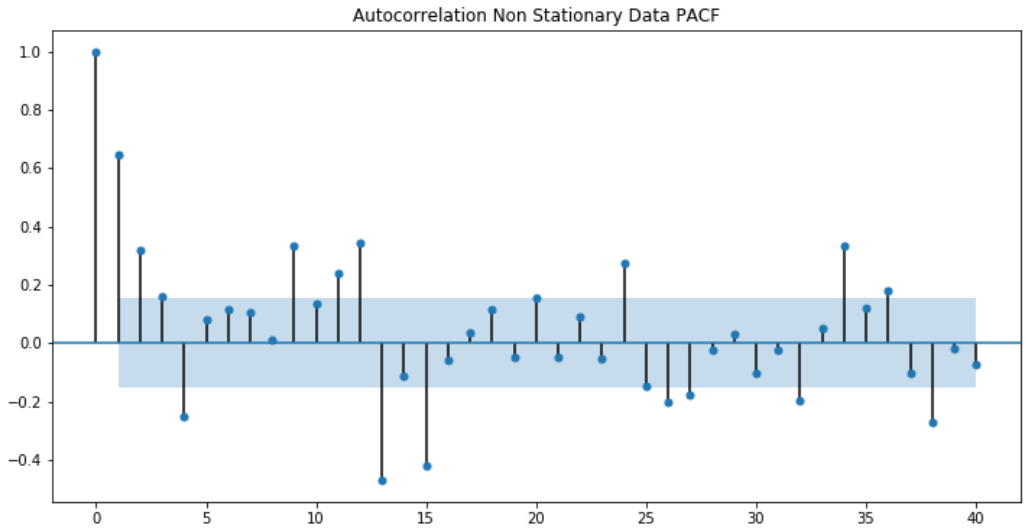



Рисунок 3.4 – Часткова автокореляційна функція народжуваності України

Із побудованих функцій з використанням 40 лагів, можна зробити висновки, що побудову моделей необхідно здійснювати з першого, другого та тринадцятого порядку, адже скоріш за все вони мають найбільш адекватну оцінку моделювання та кращу якість прогнозу.

Розпочнемо із авторегресійної моделі першого порядку, яка матиме наступний вигляд:

$$y(k) = a_0 + a_1 y(k-1) \ + \varepsilon(k) = 3.7381 + 0.621858\, y(k-1) + \varepsilon(k)$$

Модель AR(2) має вигляд:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \varepsilon(k) =$$
$$= 2.523 + 0.410139\, y(k-1) + 0.325851\, y(k-2) + \varepsilon(k)$$

Модель AR(13) має вигляд:

$$\begin{aligned} y(k) = a_0 &+ a_1 y(k-1) + a_2 y(k-2) + a_3 y(k-3) + a_4 y(k-4) \\ &+ a_5 y(k-5) + a_6 y(k-6) + a_7 y(k-7) + a_8 y(k-8) \\ &+ a_9 y(k-9) + a_{10} y(k-10) + a_{11} y(k-11) + a_{12} y(k-12) \\ &+ a_{13} y(k-13) + \varepsilon(k) = \\ 2.012 &+ 0.3734\, y(k-1) + 0.245937\, y(k-2) + 0.168274\, y(k-3) \\ &- 0.171265\, y(k-4) - 0.005166\, y(k-5) + 0.28109\, y(k-6) \\ &+ 0.019134\, y(k-7) - 0.100341\, y(k-8) + 0.011192\, y(k-9) \\ &+ 0.031772\, y(k-10) + 0.201880\, y(k-11) \\ &+ 0.501992\, y(k-12) - 0.408821\, y(k-13) + \varepsilon(k) \end{aligned}$$

Тепер зобразимо отримані моделі AR(1), AR(2), AR(13) на порівняльному Рисунку 3.5.

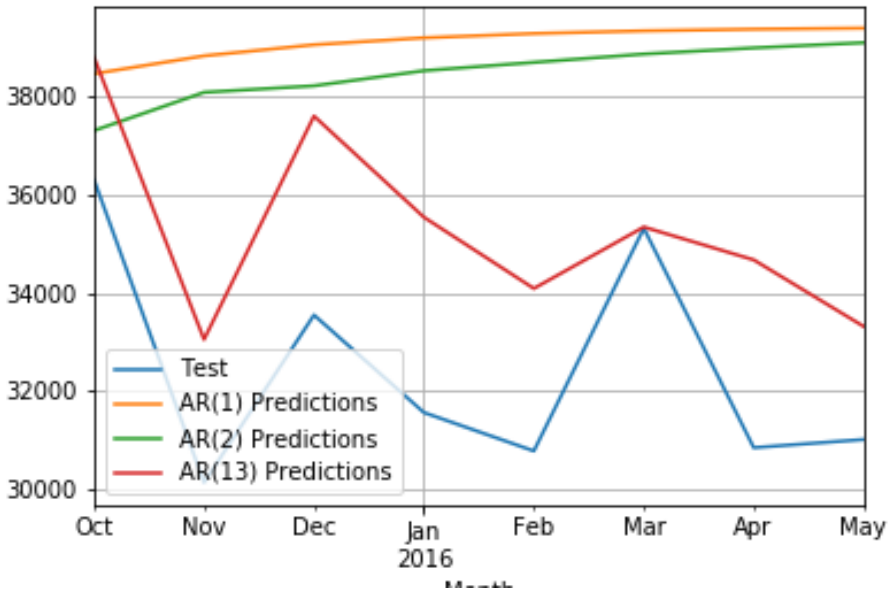



Рисунок 3.5 – Порівняльний графік моделей AR(1), AR(2), AR(13)

Даний графік демонструє актуальність побудови моделей в порівнянні з реальними тестовими даними(позначені синім) за період жовтень 2015 р. по травень 2016 р. На ньому можемо спостерігати, що найкраще відтворює динаміку зміни народжуваності саме модель AR(13).

Далі врахуємо тренди різних порядків та зобразимо їх на тій же самій тестовій вибірці за період жовтень 2015 р. по травень 2016 р на Рисунку 3.6.

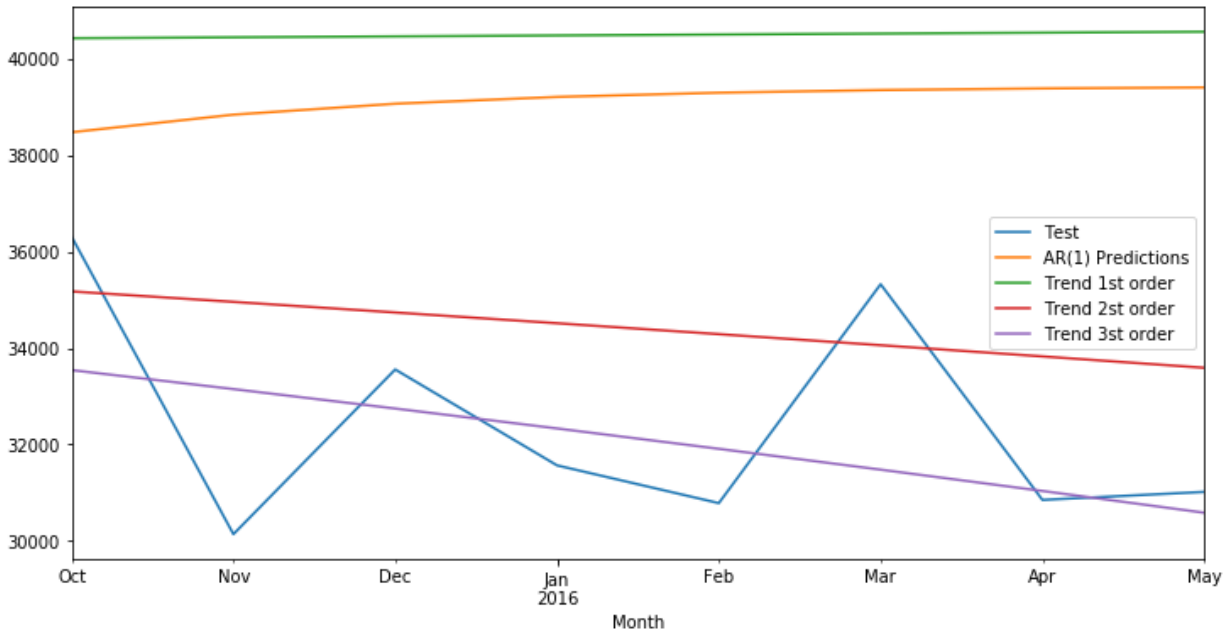



Рисунок 3.6 – Візуалізація трендів різних порядків

Перейдемо до авторегресійної моделі з ковзним середнім, а саме до моделі ARMA(13,1) що має наступний вигляд:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + a_3 y(k-3) + a_4 y(k-4) + a_5 y(k-5) + a_6 y(k-6) + a_7 y(k-7) + a_8 y(k-8) + a_9 y(k-9) + a_{10} y(k-10) + a_{11} y(k-11) + a_{12} y(k-12) + a_{13} y(k-13) + \varepsilon(k) + \beta_1 \varepsilon(k-1) =$$

$$3.804 + 0.8681\, y(k-1) + 0.1677\, y(k-2) + 0.0505\, y(k-3) - 0.2607\, y(k-4) + 0.1010\, y(k-5) + 0.0152\, y(k-6) + 0.0182\, y(k-7) - 0.1102\, y(k-8) + 0.0806\, y(k-9) - 0.0091\, y(k-10) + 0.2050\, y(k-11) + 0.4298\, y(k-12) - 0.5887\, y(k-13) + \varepsilon(k) - 0.6443\, \varepsilon(k-1)$$

Візуалізуємо отриману модель на графіку в порівнянні з тестовою вибіркою за період жовтень 2015 р. по травень 2016 р на Рисунку 3.7.

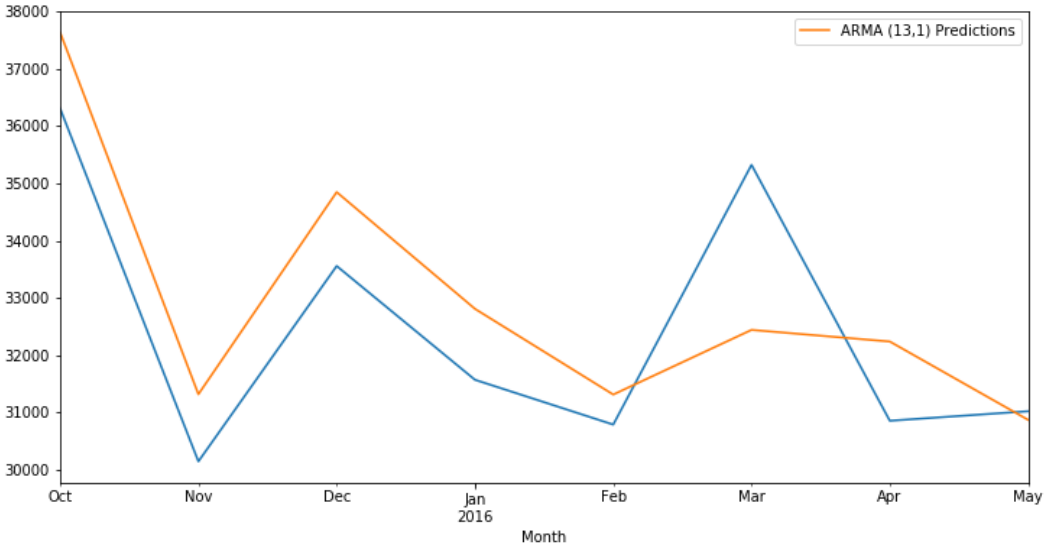



Рисунок 3.7 – Візуалізація моделі ARMA(13,1)

Побудуємо модель ARIMA(13,1,1), яка зображена на Рисунку 3.8 .

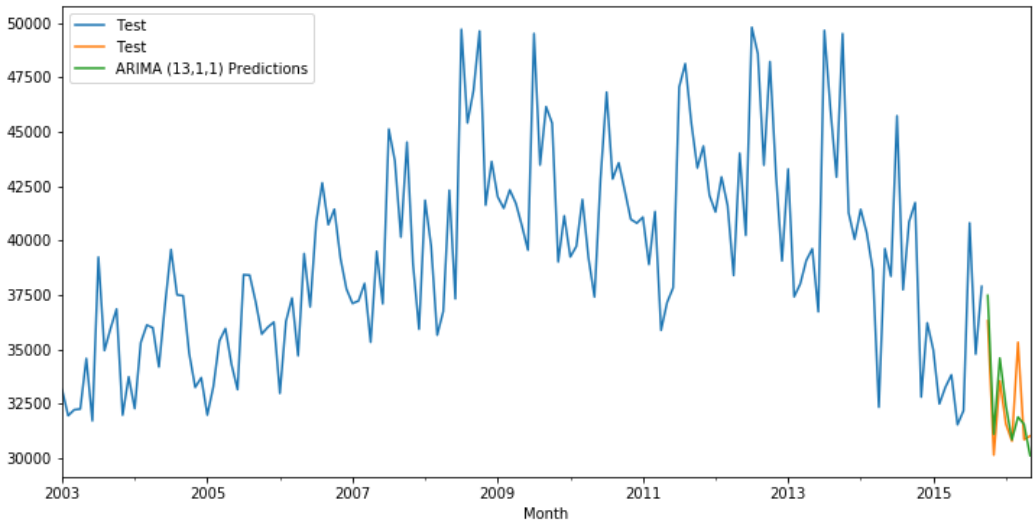



Рисунок 3.8 – Візуалізація моделі ARIMA(13,1,1)

Оскілки всі моделі, що цікавили, були розглянуті, можна оцінити їхні критерії адекватності та характеристи якості прогнозу. Ці дані, зібрані в таблицю 3.1 для наочної демонстрації, дають можливість якісно оцінити адекватність моделі та її прогнозу, а також встановити доцільність практичного застосування отриманого прогнозу.

Таблиця 3.1 – Висновки, отримані на основі побудованих моделей

| Тип моделі | Критерії адекватності | | | Оцінки прогнозу | | |
|---|---|---|---|---|---|---|
| | $R^2$ | $\sum e^2(k)$ | $DW$ | СеКП | САПП | Theil |
| AR(1) | 0.395 | 1.112 | 2.398 | 0.079 | 0.631 | 0.0039 |
| AR(2) | 0.408 | 1.067 | 2.023 | 0.084 | 0.601 | 0.0037 |
| AR(13) | 0.607 | 0.889 | 1.859 | 0.078 | 0.571 | 0.0034 |
| ARMA(13,1) | 0.695 | 0.493 | 1.989 | 0.059 | 0.479 | 0.0029 |
| ARIMA(13,1,1) | 0.768 | 0.345 | 1.969 | 0.048 | 0.363 | 0.0023 |

За продемонстрованими результатами можна зробити висновок, що поміж всіх моделей, що були побудовані, зважаючи на критерії адекватності модель ARIMA(13,1,1) має найкращі показники.

Нарешті можна продемонструвати прогноз народжуваності України на 3 роки вперед, що візуалізований на Рисунку 3.9.

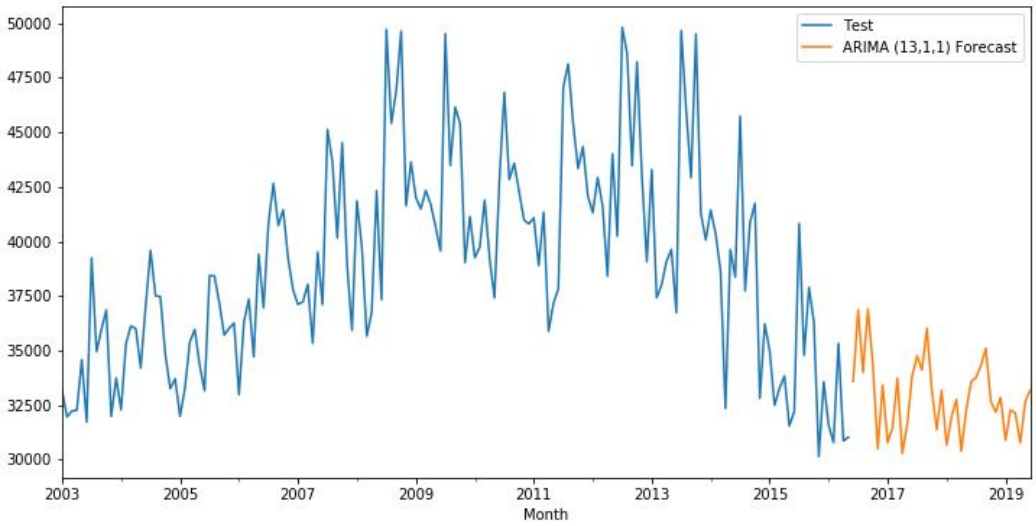



Рисунок 3.9 – Прогноз народжуваності України на 3 роки.

### 3.3.2 Регресійні моделі та прогноз кількості населення України

Однією з основних характеристик демографічного стану держави є її чисельність населення, тому в наступному підрозділі буде розглянуто особливості поведінки даного показника вибірка якого налічує 161 елемент помісячно на проміжку 2003-2016 років.

Зобразимо стан чисельності населення за період 2003-2016 рр. на Рисунку 3.10:

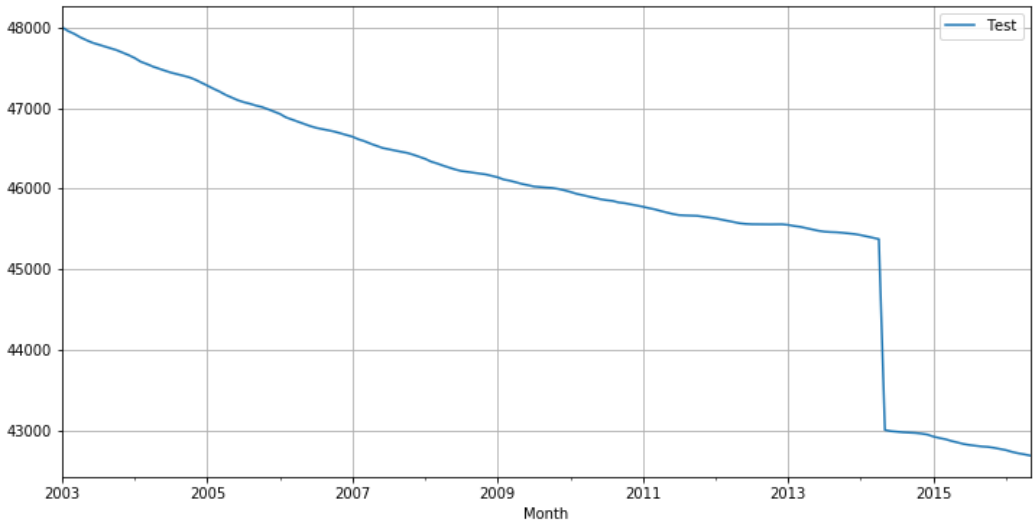



Рисунок 3.10 – Реальний стан чисельності за період 2003-2016 рр.

Для кращого вибору порядку регресійних моделей побудуємо ACF(АКФ) - автокореляційну функцію та PACF(ЧАКФ) – часткову автокореляційну функцію на наступних рисунках відповідно Рисунок 3.11 та Рисунок 3.12.

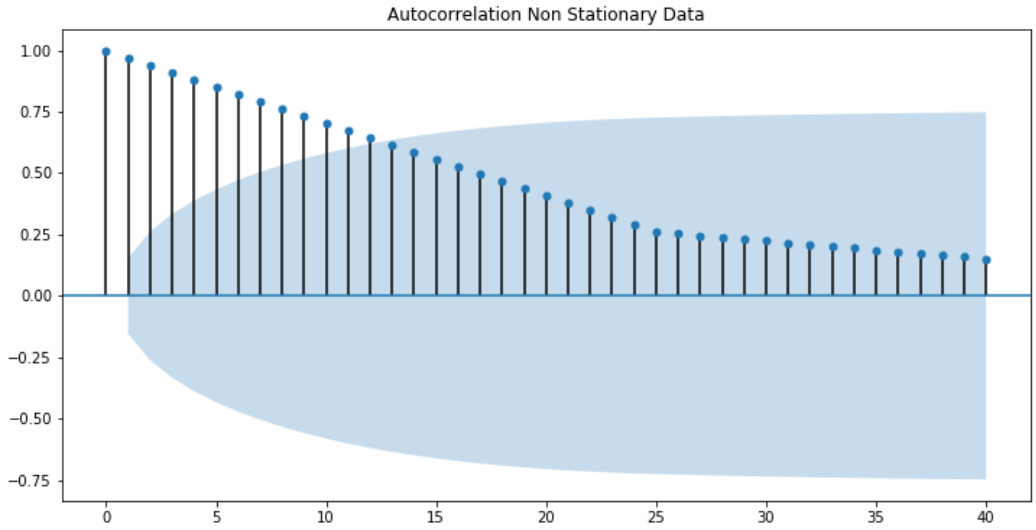



Рисунок 3.11 – Автокореляційна функція чисельності України

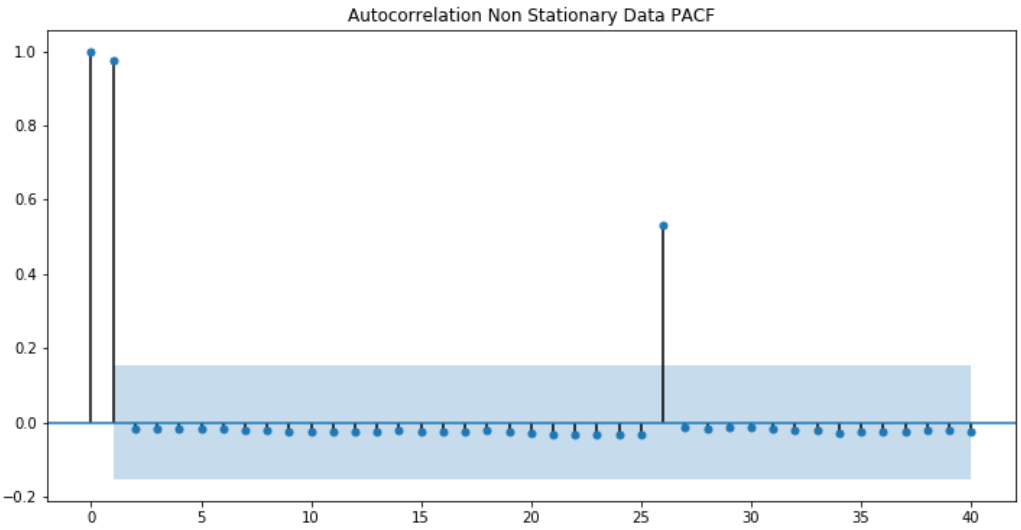



Рисунок 3.12 – Часткова автокореляційна функція чисельності України

Із побудованих функцій з використанням 40 лагів, можна зробити висновки, що побудову моделей необхідно здійснювати з першого порядку, адже скоріш за все він має найбільш адекватну оцінку моделювання та кращу якість прогнозу.

Модель AR(1):

$$y(k) = a_0 + a_1 y(k-1) \ + \varepsilon(k) = -0.0104 + 1.001535\, y(k-1) + \varepsilon(k)$$

Дана модель має наступні критерії адекватності моделі:

$$R^2 = 0.9821;\ \sum e^2(k) = 0.0027;\ DW = 2.0098$$

Де $R^2$ прямує до одиниці, сума квадратів похибок близька нулю, а $DW$ до двійки, що є показником ефективності використання даної моделі.

До того ж дана модель має гарні показники оцінки якості прогнозу, а саме:

$$\text{СеКП} = 0.0034; \text{САПП} = 0{,}027\%; \text{Coef}THeil = 0.0001$$

На наступному Рисунку 3.13 ми зможемо спостерігати короткостроковий прогноз чисельності населення України на 3 роки.

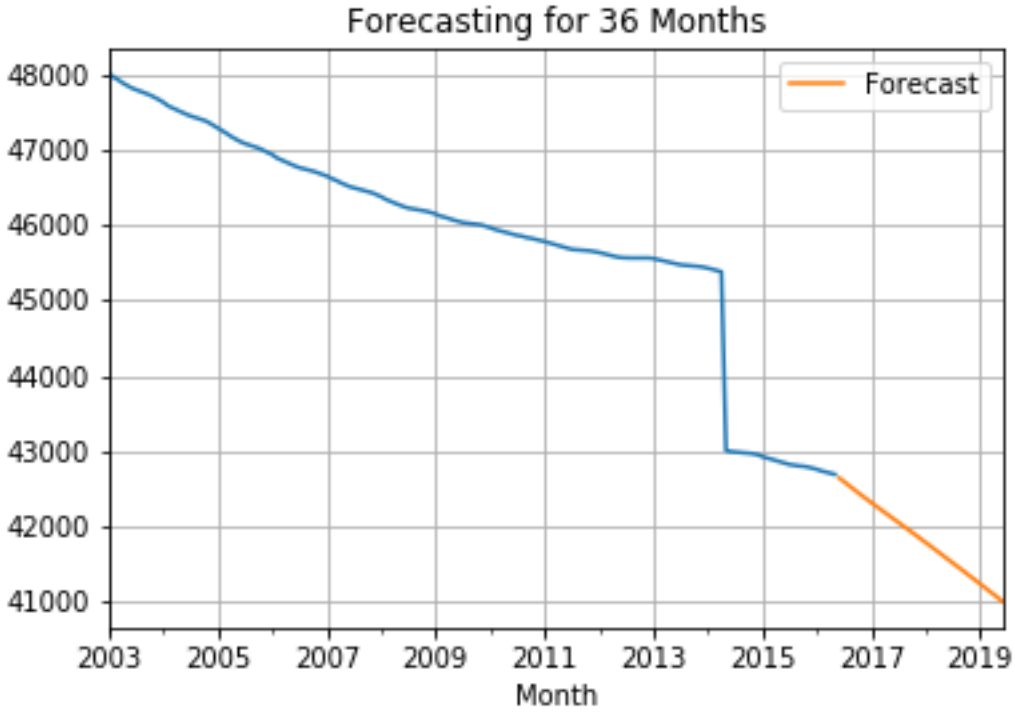



Рисунок 3.13 – Прогноз чисельності населення України на 3 роки.

### 3.3.3 Регресійні моделі та прогноз смертності України

Однією з основних характеристик демографічного стану держави є її смертності, тому в наступному підрозділі буде розглянуто особливості поведінки даного показника вибірка якого налічує 161 елемент помісячно на проміжку 2003-2016 років.

Зобразимо стан смертності за період 2003-2016 рр. на Рисунку 3.14:

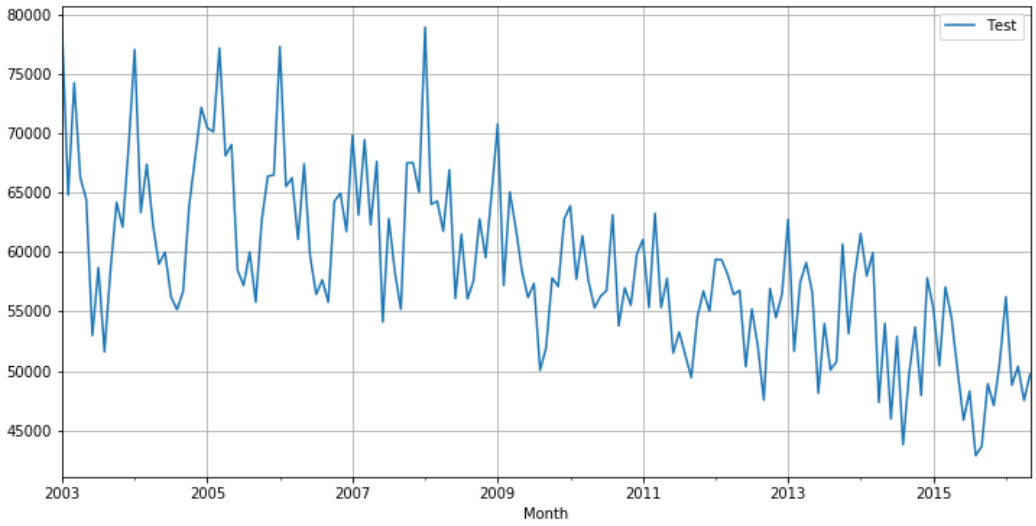



Рисунок 3.14 – Реальний стан смертності за період 2003-2016 рр.

Для кращого вибору порядку регресійних моделей побудуємо ACF(АКФ) - автокореляційну функцію та PACF(ЧАКФ) – часткову автокореляційну функцію на наступних рисунках відповідно Рисунок 3.15 та Рисунок 3.16.

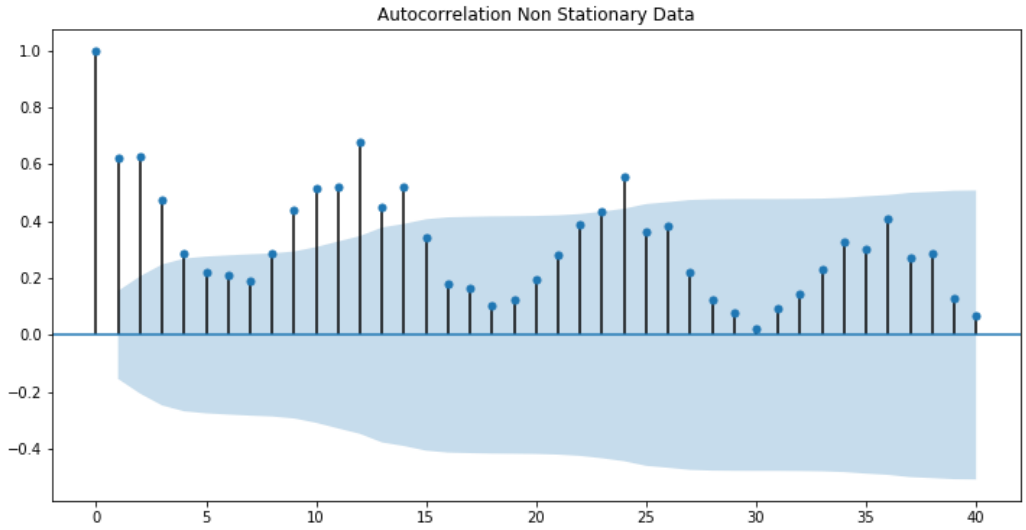



Рисунок 3.15 – Автокореляційна функція смертності України

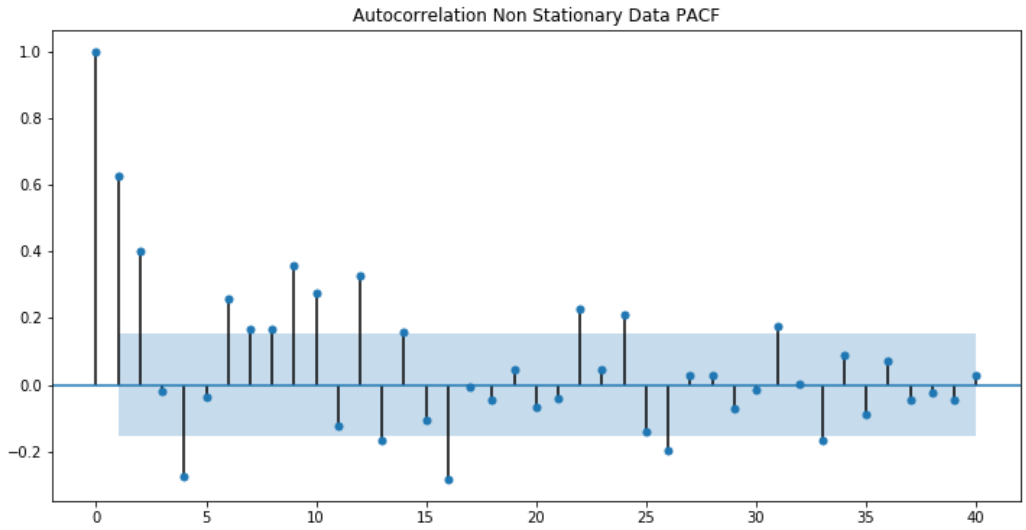



Рисунок 3.16 – Часткова автокореляційна функція смертності України

Із побудованих функцій з використанням 40 лагів, можна зробити висновки, що побудову моделей необхідно здійснювати з першого, другого та дванадцятого порядку, адже скоріш за все вони мають найбільш адекватну оцінку моделювання та кращу якість прогнозу.

Розпочнемо із авторегресійної моделі першого порядку, яка матиме наступний вигляд:

$$y(k) = a_0 + a_1 y(k-1) \ + \varepsilon(k) = 3.9705 + 0.609101\, y(k-1) + \varepsilon(k)$$

Модель AR(2) має вигляд:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \varepsilon(k) =$$
$$= 2.412 + 0.381768\, y(k-1) + 0.403037\, y(k-2) + \varepsilon(k)$$

Модель AR(12) має вигляд:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + a_3 y(k-3) + a_4 y(k-4) + a_5 y(k-5) + a_6 y(k-6) + a_7 y(k-7) + a_8 y(k-8) + a_9 y(k-9) + a_{10} y(k-10) + a_{11} y(k-11) + a_{12} y(k-12) + a_{13} y(k-13) + \varepsilon(k) =$$

$$-1{,}197 + 0.094061\, y(k-1) + 0.340682\, y(k-2) + 0.132125\, y(k-3) - 0.135496\, y(k-4) + 0.026736\, y(k-5) + 0.002574\, y(k-6) - 0.161476\, y(k-7) - 0.020313\, y(k-8) + 0.268825\, y(k-9) + 0.183726\, y(k-10) - 0.071671\, y(k-11) + 0.392301\, y(k-12) + \varepsilon(k)$$

Тепер зобразимо отримані моделі AR(1), AR(2), AR(12) на порівняльному Рисунку 3.17.


Рисунок 3.17 – Порівняльний графік моделей AR(1), AR(2), AR(12)

Даний графік демонструє актуальність побудови моделей в порівнянні з реальними тестовими даними(позначені синім) за період жовтень 2015 р. по травень 2016 р. На ньому можемо спостерігати, що найкраще відтворює динаміку зміни народжуваності саме модель AR(12).

Побудуємо модель ARIMA(12,1,1), яка зображена на Рисунку 3.18 .

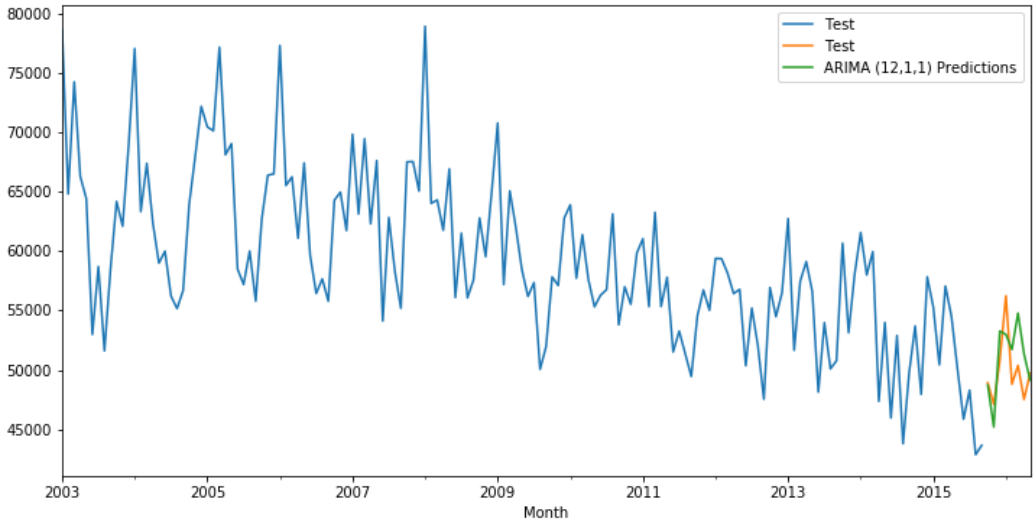



Рисунок 3.18 – Візуалізація моделі ARIMA(12,1,1)

Оскілки всі моделі, що цікавили, були розглянуті, можна оцінити їхні критерії адекватності та характеристи якості прогнозу. Ці дані, зібрані в таблицю 3.2 для наочної демонстрації, дають можливість якісно оцінити адекватність моделі та її прогнозу, а також встановити доцільність практичного застосування отриманого прогнозу.

Таблиця 3.2 – Висновки, отримані на основі побудованих моделей

| Тип моделі | Критерії адекватності | | | Оцінки прогнозу | | |
|---|---|---|---|---|---|---|
| | $R^2$ | $\sum e^2(k)$ | $DW$ | СеКП | САПП | Theil |
| AR(1) | 0.397 | 1.298 | 2.495 | 0.090 | 0.668 | 0.0038 |
| AR(2) | 0.505 | 1.089 | 2.003 | 0.082 | 0.606 | 0.0038 |
| AR(12) | 0.759 | 0.501 | 1.937 | 0.061 | 0.459 | 0.0030 |
| ARIMA(12,1,1) | 0.768 | 0.498 | 1.945 | 0.055 | 0.419 | 0.0025 |

За продемонстрованими результатами можна зробити висновок, що поміж всіх моделей, що були побудовані, зважаючи на критерії адекватності модель ARIMA(12,1,1) має найкращі показники.

Отже, можна продемонструвати прогноз смертності України на 3 роки вперед, що візуалізований на Рисунку 3.19.

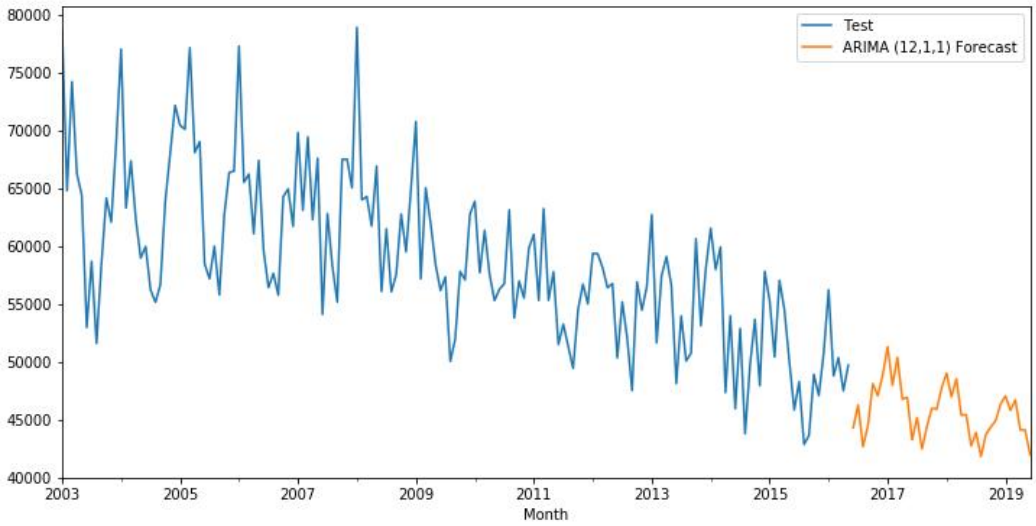



Рисунок 3.19 – Прогноз смертності України на 3 роки.

Висновки до розділу 3

Було продемонстровано різні регресійні моделі прогнозування, які можна використовувати в демографічних процесах. Також реалізовано програмний продукт за допомогою якого недосвідчений користувач може з легкістю виконувати поставлені завдання пов’язані з дослідженням демографічних процесів, а саме, конкретно в даній роботі, характер поведінки загальної чисельності населення України та, впливаючими на неї факторами, такими як народжуваність та смертність.

Побудовано три види регресійних моделей різних порядків: AR, ARMA, ARIMA. А також виконано інші операції, які допомагають виконати якісний аналіз даних моделей. На основі отриманих моделей було обрано найбільш сприятливі для прогнозування за різними критеріями адекватності моделі, а також за допомогою оцінок якості прогнозу побудовано короткостроковий прогноз та проаналізовано його практичну цінність.

# РОЗДІЛ 4
# ФУНКЦІОНАЛЬНО-ВАРТІСНИЙ АНАЛІЗ ПРОГРАМНОГО ПРОДУКТУ

У даному розділі проводиться оцінка основних характеристик програмного продукту, розробленого в рамках дипломної роботи. Програмний продукт написаний на мові програмування Python 3.6 у середовищі розробки Jupyter Notebook.

В даному розділі проводиться аналіз варіантів реалізації модулю з метою вибору оптимального, з економічної точки зору. А саме проводиться функціонально-вартісний аналіз (ФВА).

Функціонально-вартісний аналіз — це метод комплексного техніко-економічного дослідження об’єкта з метою розвитку його корисних функцій при оптимальному співвідношенні між їхньою значимістю для споживача і витратами на їхнє здійснення.

Є одним з основних методів оцінки вартості науково-дослідної роботи, оскільки ФВА враховує як технічну оцінку продукту, що розробляється, так і економічну частину розробки.

Крім того, даний метод дозволяє вибрати оптимальний, як з погляду розробника, так і з точки зору покупця варіант розв’язання будь-якої задачі, а також дозволяє оптимізувати витрати й час виконання робіт.

Мета ФВА полягає у забезпеченні правильного розподілу ресурсів, виділених на виробництво продукції або надання послуг, на прямі та непрямі витрати. У даному випадку – аналізу функцій програмного продукту й виявлення усіх витрат на реалізацію цих функцій.

Фактично цей метод працює за таким алгоритмом:

а) визначається послідовність функцій, необхідних для виробництва продукту. Спочатку – всі можливі, потім вони розподіляються по двом групам: ті, що впливають на вартість

продукту і ті, що не впливають. На цьому ж етапі оптимізується сама послідовність скороченням кроків, що не впливають на цінність і відповідно витрат.

б) для кожної функції визначаються повні річні витрати й кількість робочих часів.

в) для кожної функції на основі оцінок попереднього пункту визначається кількісна характеристика джерел витрат.

г) після того, як для кожної функції будуть визначені їх джерела витрат, проводиться кінцевий розрахунок витрат на виробництво продукту.

## 4.1 Постановка задачі техніко-економічного аналізу

У роботі застосовується метод ФВА для проведення техніко-економічний аналізу розробки.

Відповідно цьому варто обирати і систему показників якості програмного продукту.

Технічні вимоги до продукту наступні:

– програмний продукт повинен функціонувати на персональних комп’ютерах із стандартним набором компонент;

– забезпечувати високу швидкість обробки великих об’ємів даних у реальному часі;

– забезпечувати зручність і простоту взаємодії з користувачем або з розробником програмного забезпечення у випадку використовування його як модуля;

– передбачати мінімальні витрати на впровадження програмного продукту.

### 4.1.1 Обґрунтування функцій програмного продукту

Головна функція $F_0$ – розробка програмного продукту, який аналізує процес за вхідними даними та будує його модель для подальшого прогнозування. Виходячи з конкретної мети, можна виділити наступні основні функції ПП:

$F_1$ – вибір мови програмування;

$F_2$ – вибір оптимального середовища розробки;

$F_3$ – інтерфейс користувача.

Кожна з основних функцій може мати декілька варіантів реалізації.

Функція $F_1$:

а) мова програмування Eviews;

б) мова програмування Python.

Функція $F_2$:

а) IDE;

б) Jupyter Notebook.

Функція $F_3$:

а) додаток в браузері;

б) консольний додаток.

### 4.1.2 Варіанти реалізації основних функцій

Варіанти реалізації основних функцій наведені у морфологічній карті системи (рис. 4.1). На основі цієї карти побудовано позитивно-негативну матрицю варіантів основних функцій (таблиця 4.1).

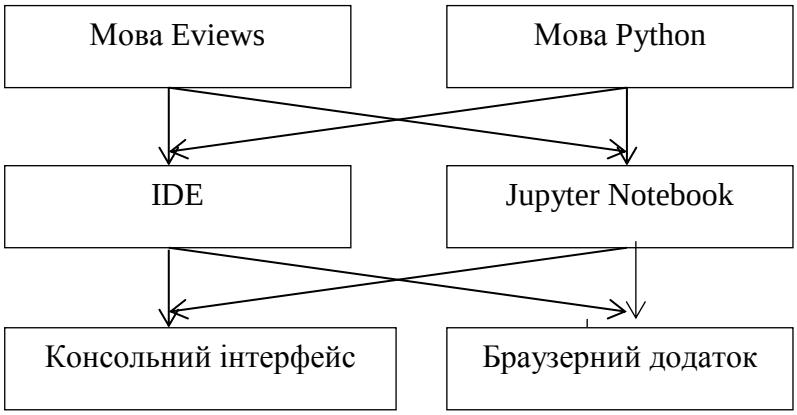



Рисунок 4.1 – Морфологічна карта

Морфологічна карта відображує всі можливі комбінації варіантів реалізації функцій, які складають повну множину варіантів ПП.

Таблиця 4.1 – Позитивно-негативна матриця

| Основні функції | Варіанти реалізації | Переваги | Недоліки |
|---|---|---|---|
| *F1* | *А* | Повністю готове середовище для роботи | Не має безкоштовної ліцензії |
| | *Б* | Простота реалізації в рамках даної задачі | Швидкодія |
| *F2* | А | Виділення синтаксису | Довгий процес виправлення помилок |
| | Б | Можливість виконувати програму блоками крок за кроком. | Часозатратний |
| *F3* | *А* | Не вимогливий до потужностей комп’ютера | Не 'user-friendly' |
| | *Б* | Зручність для пересічного користувача | Повільніше працює |

На основі аналізу позитивно-негативної матриці робимо висновок, що при розробці програмного продукту деякі варіанти реалізації функцій варто відкинути, тому, що вони не відповідають поставленим перед програмним продуктом задачам. Ці варіанти відзначені у морфологічній карті.

Функція *F1:*

Оскільки в рамках даної задачі налаштування Python не займають багато часу, обираємо варіант Б.

Функція *F2:*

При роботі з даними необхідно постійно виправляти та доповнювати код, тому вибираємо варіант Б.

Функція *F3:*

Оскільки, щодо інтерфейсу програмного продукту не має вишуканих вимог (ПП є виключно робочою частиною, в подальшому планується абсолютна автоматизація), то обидва варіанти А і Б влаштовують.

Таким чином, будемо розглядати такі варіанти реалізації ПП:

1. F1Б – F2Б – F3А
2. F1Б – F2Б – F3Б

Для оцінювання якості розглянутих функцій обрана система параметрів, описана нижче.

## 4.2 Обґрунтування системи параметрів ПП

### 4.2.1 Опис параметрів

На підставі даних про основні функції, що повинен реалізувати програмний продукт, вимог до нього, визначаються основні параметри виробу, що будуть використані для розрахунку коефіцієнта технічного рівня.

Для того, щоб охарактеризувати програмний продукт, будемо використовувати наступні параметри:

– *X1* – швидкодія мови програмування;

– *X2* – час обробки даних;

– *X3* – потенційний об’єм програмного коду.

*X1*: Відображає швидкодію операцій залежно від обраної мови програмування.

*X2:* Відображає час, який витрачається на дії.

*X3:* Показує розмір програмного коду який необхідно створити безпосередньо розробнику.

### 4.2.2 Кількісна оцінка параметрів

Гірші, середні і кращі значення параметрів вибираються на основі вимог замовника й умов, що характеризують експлуатацію ПП як показано у табл. 4.2.

Таблиця 4.2 – Основні параметри ПП

| Назва Параметра | Умовні позначення | Одиниці виміру | Значення параметра | | |
|---|---|---|---|---|---|
| | | | гірші | середні | кращі |
| Швидкодія мови програмування | X1 | Оп/мс | 19000 | 11000 | 2000 |
| Час обробки запитів | X2 | мс | 1000 | 420 | 60 |
| Потенційний об’єм прогр. коду | X3 | кількість строк коду | 350 | 200 | 100 |

За даними таблиці 4.2 будуються графічні характеристики параметрів – рис. 4.2 – рис. 4.4.

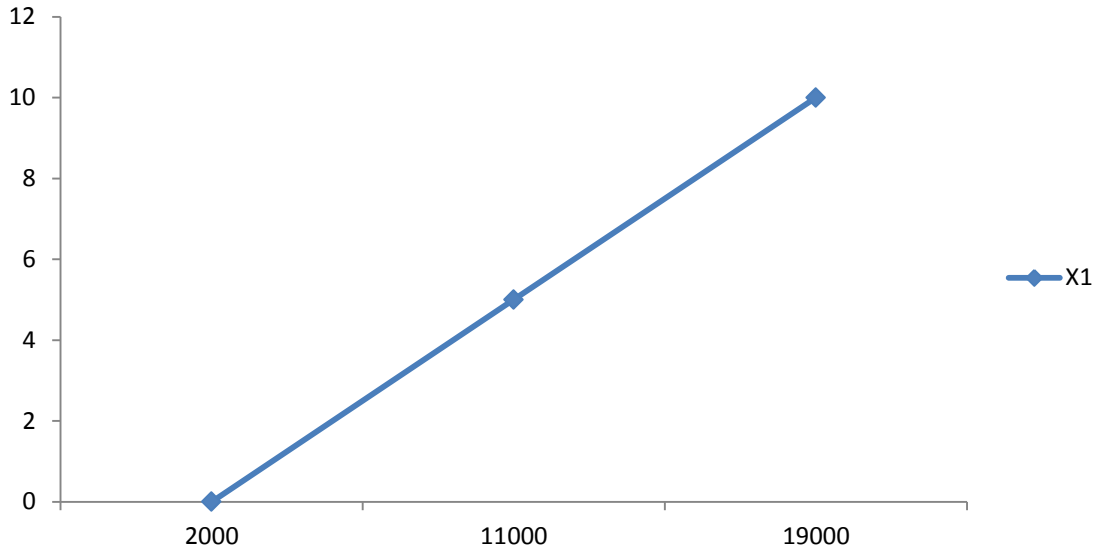



Рисунок 4.2 – X1, швидкодія мови програмування

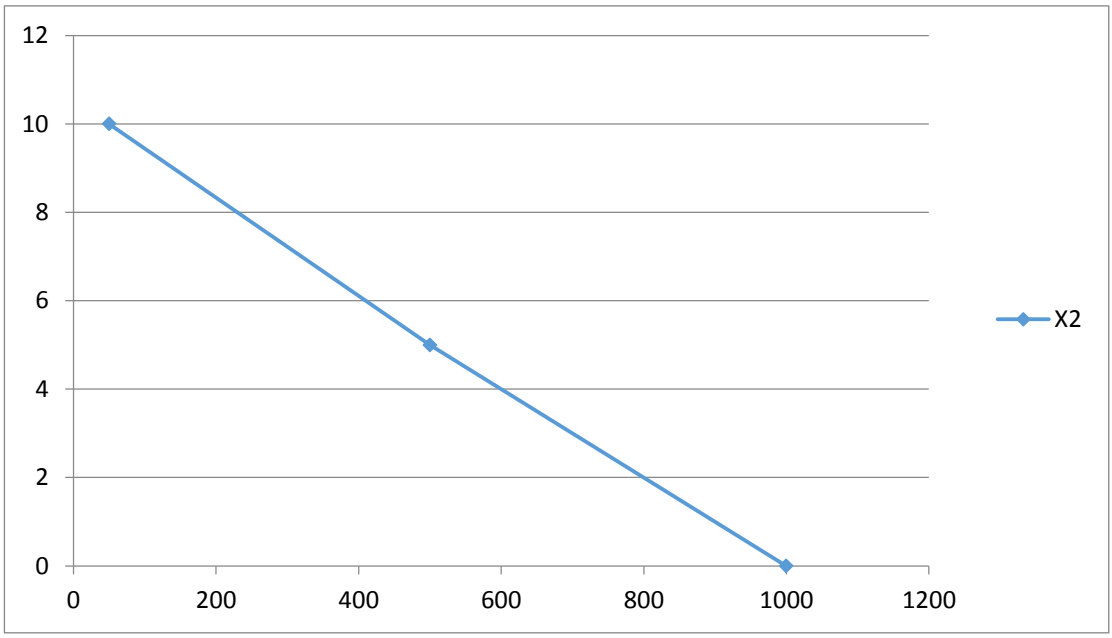



Рисунок 4.3 – X2, час виконання запитів користувача

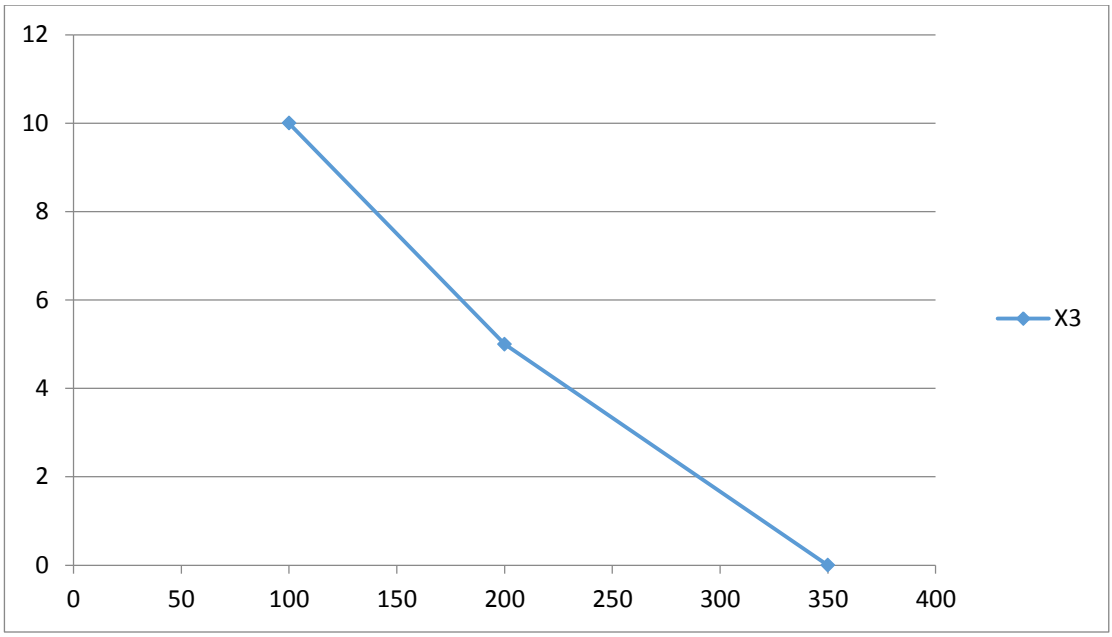



Рисунок 4.4 – X3, потенційний об’єм програмного коду

### 4.2.3 Аналіз експертного оцінювання параметрів

Після детального обговорення й аналізу кожний експерт оцінює ступінь важливості кожного параметру для конкретно поставленої цілі – розробка програмного продукту, який дає найбільш точні результати при знаходженні параметрів моделей адаптивного прогнозування і обчислення прогнозних значень.

Значимість кожного параметра визначається методом попарного порівняння. Оцінку проводить експертна комісія із 7 людей. Визначення коефіцієнтів значимості передбачає:

– визначення рівня значимості параметра шляхом присвоєння різних рангів;

– перевірку придатності експертних оцінок для подальшого використання;

– визначення оцінки попарного пріоритету параметрів;

– обробку результатів та визначення коефіцієнту значимості.

Результати експертного ранжування наведені у таблиці 4.3.

Таблиця 4.3 – Результати ранжування параметрів

| Позначення параметра | Назва параметра | Одиниці виміру | Ранг параметра за оцінкою експерта | | | | | | | Сума рангів $R_i$ | Відхилення $\Delta_i$ | $\Delta_i^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | | |
| *X1* | Швидкодія мови програмування | Оп/мс | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 12 | 0,33 | 0,109 |
| *X2* | Час обробки запитів | Мс | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 12 | -14,67 | 215,2 |
| *X3* | Потенційний об’єм програмного коду | кількість строк коду | 2 | 2 | 3 | 3 | 3 | 2 | 3 | 18 | 14,33 | 214,92 |
| | Разом | | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 42 | 0 | 430,229 |

Для перевірки степені достовірності експертних оцінок, визначимо наступні параметри:

а) сума рангів кожного з параметрів і загальна сума рангів:

$$R_i = \sum_{j=1}^{N} r_{ij} R_{ij} = \frac{Nn(n+1)}{2} = 80, \quad (4.1)$$

де $N$ – число експертів, $n$ – кількість параметрів;

б) середня сума рангів:

$$T = \frac{1}{n} R_{ij} = 26{,}67 \quad (4.2)$$

в) відхилення суми рангів кожного параметра від середньої суми рангів:

$$\Delta_i = R_i - T \quad (4.3)$$

Сума відхилень по всім параметрам повинна дорівнювати 0;

г)загальна сума квадратів відхилення:

$$S = \sum_{i=1}^{N} \Delta_i^2 = 430{,}229 \quad (4.4)$$

Порахуємо коефіцієнт узгодженості:

$$W = \frac{12S}{N^2(n^3-n)} = \frac{12 \cdot 430{,}229}{7^2(3^3-3)} = 4{,}39 > W_{\text{норм}} = 0{,}67 \qquad (4.5)$$

Ранжування можна вважати достовірним, тому що знайдений коефіцієнт узгодженості перевищує нормативний, котрий дорівнює 0,67.

Скориставшись результатами ранжирування, проведемо попарне порівняння всіх параметрів і результати занесемо у таблицю 4.4.

Таблиця 4.4 – Попарне порівняння параметрів

| Параметри | Експерти | | | | | | | Кінцева оцінка | Числове значення |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | |
| X1 і X2 | > | > | > | > | > | > | > | > | 1,5 |
| X1 і X3 | < | < | < | < | < | < | < | < | 0,5 |
| X2 і X3 | < | < | < | < | < | < | < | < | 0,5 |

Числове значення, що визначає ступінь переваги *i*–го параметра над *j*–тим, $a_{ij}$ визначається по формулі:

$$a_{ij} = \begin{cases} 1.5, \text{при Xi} > \text{Xj} \\ 1.0, \text{при Xi} = \text{Xj} \\ 0.5, \text{при Xi} < \text{Xj} \end{cases} \qquad (4.6)$$

З отриманих числових оцінок переваги складемо матрицю A=$\| a_{ij} \|$ .Для кожного параметра зробимо розрахунок вагомості $K_{ві}$ за наступними формулами:

$$K_{\text{ві}} = \frac{b_i}{\sum_{i=1}^{n} b_i}, \quad (4.7)$$

де $b_i = \sum_{i=1}^{N} a_{ij}$.

Відносні оцінки розраховуються декілька разів доти, поки наступні значення не будуть незначно відрізнятися від попередніх (менше 2%).

На другому і наступних кроках відносні оцінки розраховуються за наступними формулами:

$$K_{\text{ві}} = \frac{b_i'}{\sum_{i=1}^{n} b_i'}, \text{де } b_i' = \sum_{i=1}^{N} a_{ij} b_j. \quad (4.8)$$

Як видно з таблиці 5.5, різниця значень коефіцієнтів вагомості не перевищує 2%, тому більшої кількості ітерацій не потрібно

Таблиця 4.5 – Розрахунок вагомості параметрів

| Параметри $x_i$ | Параметри $x_j$ | | | Перша ітер. | | Друга ітер. | | Третя ітер. | |
|---|---|---|---|---|---|---|---|---|---|
| | X1 | X2 | X3 | $b_i$ | $K_{\text{ві}}$ | $b_i^1$ | $K_{\text{ві}}^1$ | $b_i^2$ | $K_{\text{ві}}^2$ |
| X1 | 1,0 | 1,5 | 0,5 | 3.0 | 0.333 | 8.0 | 0.32 | 21.25 | 0.31 |
| X2 | 0,5 | 1,0 | 0,5 | 2.0 | 0.222 | 5.5 | 0.22 | 15.25 | 0.223 |
| X3 | 1,5 | 1,5 | 1,0 | 4.0 | 0. 445 | 11.5 | 0.46 | 31.75 | 0.465 |
| Всього: | | | | 9 | 1 | 25 | 1 | 68.25 | 1 |

## 4.3 Аналіз рівня якості варіантів реалізації функцій

Визначаємо рівень якості кожного варіанту виконання основних функцій окремо. Коефіцієнт технічного рівня для кожного варіанта реалізації ПП розраховується так (таблиця 4.6):

$$K_K(j) = \sum_{i=1}^{n} K_{вi,j} B_{i,j}, \quad (4.9)$$

де $n$ – кількість параметрів;

$K_{вi}$– коефіцієнт вагомості $i$–го параметра;

$B_i$ – оцінка $i$–го параметра в балах.

Таблиця 4.6 – Розрахунок показників рівня якості варіантів реалізації основних функцій ПП

| Основні функції | Варіант реалізації функції | Параметри $x_i$ | Абсолютне значення параметра | Бальна оцінка параметра | Коефіцієнт вагомості параметра | Коефіцієнт рівня якості |
|---|---|---|---|---|---|---|
| F1 | б) | Xl | 11000 | 7 | 0,312 | 2,184 |
| | | X3 | 200 | 5 | 0,465 | 2,325 |
| F2 | б) | X2 | 100 | 3 | 0,223 | 0,669 |
| | | X3 | 200 | 7,5 | 0,465 | 3,488 |
| F3 | б) | X2 | 100 | 3,5 | 0,223 | 0,78 |
| | | X3 | 350 | 1 | 0,465 | 0,465 |
| | а) | X3 | 200 | 1 | 0,465 | 0,465 |

За даними з таблиці 4.6 за формулою

$$K_K = K_{\text{ТУ}}[F_{1k}] + K_{\text{ТУ}}[F_{2k}] + \ldots + K_{\text{ТУ}}[F_{zk}], \quad (4.10)$$

визначаємо рівень якості кожного з варіантів:

$$K_{K1} = 2{,}184+2{,}325+0{,}669+3{,}488+0{,}78+0{,}465 = 9{,}911$$

$$K_{K2} = 2{,}184+2{,}325+0{,}669+3{,}488+0{,}465 = 9{,}131$$

Як видно з розрахунків, кращим є перший варіант, для якого коефіцієнт технічного рівня має найбільше значення.

## 4.4 Економічний аналіз варіантів розробки ПП

Для визначення вартості розробки ПП спочатку проведемо розрахунок трудомісткості.

Всі варіанти включають в себе два окремих завдання:

1. Розробка проекту програмного продукту;

2. Розробка програмної оболонки;

Але варіант II реалізації програмного забезпечення включає ще одне завдання:

3. Написання алгоритму збереження інформації у вигляді компонента, зручного для візуалізації.

Завдання 1 за ступенем новизни відноситься до групи А, завдання 2 – до групи Б, завдання 3 до группи Г. За складністю алгоритми, які використовуються в завданні 1 належать до групи 1; а в завданні 2 – до групи 3. Завдання 3 відноситься за складністю до групи 3.

Для реалізації завдання 1 використовується довідкова інформація, а завдання 2 використовує інформацію у вигляді даних.

Проведемо розрахунок норм часу на розробку та програмування для кожного з завдань.

Проведемо розрахунок норм часу на розробку та програмування для кожного з завдань. Загальна трудомісткість обчислюється як

$$T_O = T_P \cdot K_П \cdot K_{СК} \cdot K_M \cdot K_{СТ} \cdot K_{СТ.М}, \quad (4.11)$$

де $T_P$ – трудомісткість розробки ПП;

$K_П$ – поправочний коефіцієнт;

$K_{СК}$ – коефіцієнт на складність вхідної інформації;

$K_M$ – коефіцієнт рівня мови програмування;

$K_{СТ}$ – коефіцієнт використання стандартних модулів і прикладних програм;

$K_{СТ.М}$ – коефіцієнт стандартного математичного забезпечення

Для першого завдання, виходячи із норм часу для завдань розрахункового характеру степеню новизни А та групи складності алгоритму 1, трудомісткість дорівнює: $T_P$ =90 людино-днів. Поправочний коефіцієнт, який враховує вид нормативно-довідкової інформації для першого завдання: $K_П$ = 1.7. Поправочний коефіцієнт, який враховує складність контролю вхідної та вихідної інформації для завдань рівний 1: $K_{СК}$ = 1. Оскільки при розробці першого завдання використовуються стандартні модулі, врахуємо

це за допомогою коефіцієнта $К_{СТ}$ = 0.8. Тоді, за формулою (4.11), загальна трудомісткість програмування першого завдання дорівнює:

$$Т_1 = 90 \cdot 1.7 \cdot 0.8 = 122.4 \text{ людино-днів.}$$

Проведемо аналогічні розрахунки для подальших завдань.

Для другого завдання (використовується алгоритм третьої групи складності, степінь новизни Б), тобто $Т_Р$ =27 людино-днів, $К_П$ =0.9, $К_{СК}$ = 1, $К_{СТ}$ =0.8:

$$Т_2 = 27 * 0.9 * 0.8 = 19.44 \text{ людино-днів.}$$

Для третього завдання (використовується алгоритм третьої групи складності, ступінь новизни Г):

$$Т_р = 15 \text{ людино-днів;} \quad К_п = 0,6; \quad К_{ст} = 1; \quad Т_З = 15 * 0,6 * 1 = 9.$$

Складаємо трудомісткість відповідних завдань для кожного з обраних варіантів реалізації програми, щоб отримати їх трудомісткість:

$$Т_I = (122.4 + 19.44) * 15 = 2127,6 \text{ людино-годин;}$$

$$Т_{II} = (122.4 + 19.44 + 9) * 15 = 2262,6 \text{ людино-годин;}$$

Найбільш високу трудомісткість має варіант II.

В розробці беруть участь два програмісти з окладом 14000 грн., один спеціаліст по цифровій обробці сигналів з окладом 22000грн. Визначимо зарплату за годину за формулою:

$$СЧ = \frac{М}{T_m \cdot t} \text{ грн.,} \qquad (4.12)$$

де М – місячний оклад працівників;

$T_m$ – кількість робочих днів тиждень;

$t$ – кількість робочих годин в день.

$$\text{СЧ} = \frac{14000 + 14000 + 22000}{3 \cdot 21 \cdot 15} = 52{,}91 \text{ грн.}$$

Тоді, розрахуємо заробітну плату за формулою

$$\text{СЗП} = \text{С}_\text{ч} \cdot \text{Т}_i \cdot \text{КД}, \qquad (4.13)$$

де $\text{С}_\text{ч}$ – величина погодинної оплати праці програміста;

$\text{Т}_i$ – трудомісткість відповідного завдання;

$\text{К}_\text{Д}$ – норматив, який враховує додаткову заробітну плату.

Зарплата розробників за варіантами становить:

I. $\text{С}_\text{ЗП} = 52{,}91 \cdot 2127{,}6 \cdot 1.2 = 135085{,}58$ грн.

II. $\text{С}_\text{ЗП} = 52{,}91 \cdot 2262{,}6 \cdot 1.2 = 143657{,}00$ грн.

Відрахування на єдиний соціальний внесок в залежності від групи професійного ризику (II клас) становить 22%:

I. $\text{С}_\text{ВІД} = \text{С}_\text{ЗП} \cdot 0.3677 = 135085{,}58 \cdot 0.22 = 29718{,}83$ грн.

II. $\text{С}_\text{ВІД} = \text{С}_\text{ЗП} \cdot 0.3677 = 143657{,}00 \cdot 0.22 = 31604{,}54$ грн.

Тепер визначимо витрати на оплату однієї машино-години. ($\text{С}_\text{М}$)

Так як одна ЕОМ обслуговує одного програміста з окладом 14000 грн., з коефіцієнтом зайнятості 0,2 то для однієї машини отримаємо:

$$С_Г = 12 \cdot М \cdot К_З = 12 \cdot 12000 \cdot 0,2 = 28800 \text{ грн.}$$

З урахуванням додаткової заробітної плати:

$$С_{ЗП} = С_Г \cdot (1 + К_З) = 28800 \cdot (1 + 0.2) = 34560 \text{ грн.}$$

Відрахування на єдиний соціальний внесок:

$$С_{ВІД} = С_{ЗП} \cdot 0.3677 = 34560 \cdot 0,22 = 7603,20 \text{ грн.}$$

Амортизаційні відрахування розраховуємо при амортизації 25% та вартості ЕОМ – 8000 грн.

$$С_А = К_{ТМ} \cdot К_А \cdot Ц_{ПР} = 1.15 \cdot 0.25 \cdot 8000 = 2300 \text{ грн.},$$

де $К_{ТМ}$– коефіцієнт, який враховує витрати на транспортування та монтаж приладу у користувача;

$К_А$– річна норма амортизації;

$Ц_{ПР}$– договірна ціна приладу.

Витрати на ремонт та профілактику розраховуємо як:

$$С_Р = К_{ТМ} \cdot Ц_{ПР} \cdot К_Р = 1.15 \cdot 8000 \cdot 0.05 = 460 \text{ грн.},$$

де $К_Р$– відсоток витрат на поточні ремонти.

Ефективний годинний фонд часу ПК за рік розраховуємо за формулою:

$$Т_{ЕФ} = (Д_К - Д_В - Д_С - Д_Р) \cdot t_З \cdot К_В \qquad (4.14)$$

$$Т_{ЕФ} = (365 - 104 - 8 - 16) \cdot 8 \cdot 0.9 = 1706.4 \text{ годин},$$

де $Д_К$ – календарна кількість днів у році;

$Д_В$, $Д_С$ – відповідно кількість вихідних та святкових днів;

$Д_Р$ – кількість днів планових ремонтів устаткування;

t –кількість робочих годин в день;

$К_В$– коефіцієнт використання приладу у часі протягом зміни.

Витрати на оплату електроенергії розраховуємо за формулою:

$$С_{ЕЛ} = Т_{ЕФ} \cdot N_С \cdot К_З \cdot Ц_{ЕН} = 1706,4 \cdot 0,156 \cdot 0,9733 \cdot 2,7515 = 712,88 \text{ грн.},$$

де $N_С$ – середньо-споживча потужність приладу;

$К_З$– коефіцієнтом зайнятості приладу;

$Ц_{ЕН}$ – тариф за 1 КВт-годин електроенергії.

Накладні витрати розраховуємо за формулою:

$$С_Н = Ц_{ПР} \cdot 0.67 = 8000 \cdot 0,67 = 5360 \text{ грн.}$$

Тоді, річні експлуатаційні витрати будуть:

$$С_{ЕКС} = С_{ЗП} + С_{ВІД} + С_А + С_Р + С_{ЕЛ} + С_Н \qquad (4.15)$$

$$С_{ЕКС} = 34560 + 7603,20 + 2300 + 460 + 712,88 + 5360 = 50996,08 \text{ грн.}$$

Собівартість однієї машино-години ЕОМ дорівнюватиме:

$$С_{М\text{-}Г} = С_{ЕКС} / Т_{ЕФ} = 50996,08/1706,4 = 29,885 \text{ грн/час.}$$

Оскільки в даному випадку всі роботи, які пов‘язані з розробкою програмного продукту ведуться на ЕОМ, витрати на оплату машинного часу, в залежності від обраного варіанта реалізації, складає:

$$С_М = С_{М\text{-}Г} \cdot Т \qquad (4.15)$$

I. $С_М = 29,885 \cdot 2127,6 = 63583,326$ грн.;

II. $С_М = 29,885 \cdot 2262,6 = 67617,801$ грн.;

Накладні витрати складають 67% від заробітної плати:

$$С_Н = С_{ЗП} \cdot 0,67 \quad (4.16)$$

I. $С_Н = 135085,58 \cdot 0,67 = 90507,34$ грн.;

II. $С_Н = 143657,00 \cdot 0,67 = 96250,19$ грн.;

Отже, вартість розробки ПП за варіантами становить:

$$С_{ПП} = С_{ЗП} + С_{ВІД} + С_М + С_Н \quad (4.17)$$

I. $С_{ПП} = 135085,58 + 49670,97 + 63583,326 + 90507,34 = 338847,216$ грн.;

II. $С_{ПП} = 143657,00 + 52822,68 + 67617,801 + 96250,19 = 360347,671$ грн.;

## 4.5 Вибір кращого варіанта ПП за техніко-економічного рівня

Розрахуємо коефіцієнт техніко-економічного рівня за формулою:

$$К_{ТЕРj} = К_{Кj} / С_{Фj}, \quad (4.18)$$

$К_{ТЕР1}$ = 9,911/ 344325,79 = $2,88 \cdot 10^{-5}$;

$К_{ТЕР2}$ = 9,131/ 366173,97= $2,49 \cdot 10^{-5}$;

Як бачимо, найбільш ефективним є перший варіант реалізації програми з коефіцієнтом техніко-економічного рівня $К_{ТЕР1}$= $2,88 \cdot 10^{-5}$.

Висновки до розділу 4

В даній розрахунковій роботі проведено повний функціонально-вартісний аналіз ПП, який було розроблено в рамках дипломної роботи. Процес аналізу можна умовно розділити на дві частини.

В першій з них проведено дослідження ПП з технічної точки зору: було визначено основні функції ПП та сформовано множину варіантів їх реалізації; на основі обчислених значень параметрів, а також експертних оцінок їх важливості було обчислено коефіцієнт технічного рівня, який і дав змогу визначити оптимальну з технічної точки зору альтернативу реалізації функцій ПП.

Другу частину ФВА присвячено вибору із альтернативних варіантів реалізації найбільш економічно обґрунтованого. Порівняння запропонованих варіантів реалізації в рамках даної частини виконувалось за коефіцієнтом ефективності, для обчислення якого були обчислені такі допоміжні параметри, як трудомісткість, витрати на заробітну плату, накладні витрати.

Після виконання функціонально-вартісного аналізу програмного комплексу що розроблюється, можна зробити висновок, що з альтернатив, що залишились після першого відбору двох варіантів виконання програмного комплексу оптимальним є перший варіант реалізації програмного продукту. У нього виявився найкращий показник техніко-економічного рівня якості $К_{ТЕР}$ $2{,}88 \cdot 10^{-5}$.

Цей варіант реалізації програмного продукту має такі параметри:

- мова програмування – Python;
- використання Jupyter Notebook;
- браузерний інтерфейс.

Даний варіант виконання програмного комплексу дає користувачу відмінний функціонал, швидкодію і робить простішим виконання завдання.

# ВИСНОВКИ ПО РОБОТІ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Дана дипломна робота є результатом дослідження поведінки часових рядів та їх побудова на основі авторегресійних моделей різних типів, а також розробки програмного продукту для отримання практичних результатів та вибору найкращої моделі для наочної візуалізації даних короткострокового прогнозування демографічних процесів народонаселення України.

У роботі було досліджено та проаналізовано світові тенденції поведінки основних демографічних показників, таких як загальна чисельність, народжуваність та смертність, нестабільна поведінка яких призводить до глобальної проблеми людства. Потім була розглянута ситуація в Україні.

Під час виконання роботи було отримано низку авторегресійних моделей різних порядків. Продемонстровано порівняльний аналіз даних моделей у графічному вигляді, а також в режимі порівняльної таблиці, що сформована на основі основних критеріїв адекватності та розрахованими оцінками якості побудови прогнозу. Було зроблено висновки щодо найкращих моделей та їх актуальність у використанні певних демографічних процесів та побудовано короткостроковий прогноз. Отримані результати були апробовані на статистичних даних різних типів народонасення України.

За результатами прогнозування населення України можна зробити висновок, що тенденція зменшення чисельності населення триватиме й надалі, особливо враховуючи різкі коливання міграційних рухів. Держава в цьому випадку повинна вжити заходів для покращення умов життя середньостатистичного українця та підвищення соціально-економічного рівня громадян. Аналіз прогнозу народжуваності є також невтішним, адже кількість народжених зменшується з кожним роком, що може призвести до

старіння нації. Кількість смертей різних статево-вікових груп також має негативні наслідки, оскільки багато дітей помирає у початковому віці та чоловіків у працездатному. Останні два показника можна покращити шляхом підвищення рівня медицини в країні, соціального забезпечення молодих сімей та охорони праці.

Для вдосконалення ефективності досліджень демографічних процесів необхідно розширити область використання моделей та їхні модифікації, наприклад застосування пояснювальної змінної, що буде характеризувати вплив ВВП країни певного проміжку часу на народонаселення та інші чинники впливу. Також доцільно буде розглянути інші методи моделювання та прогнозування такі як метод подібних траєкторій або ж нейронних мереж, тощо. Збільшити кількість рівнів перевірки моделі та якості прогнозу, врахувати максимальну та мінімальну абсолютну похибку та метод максимальної правдоподібності. Покращення інтерфейсу для кращого розуміння процесу прогнозування та використання нових технологій також позитивно вплине на модернізацію даного проекту.

# ЛІТЕРАТУРА


1. Аникин А. В. Мальтус и мальтузианство / Юность науки: Жизнь и идеи мыслителей-экономистов до Маркса. 2-е изд. М. : Политиздат , 1975.266–274 с.
2. Капица С. П. Демографическая революция и будуще человечества / В мире науки. 2004. № 4. 82–91 с.
3. База даних HYDE (2016) і «Перспективи світового населення ООН» (2017). [Електронний ресурс] База даних HYDE. URL: https://themasites.pbl.nl/tridion/en/themasites/hyde/basicdrivingfactors/population/index-2.html_United Nations, Department of Economic and Social Affairs, Population Division (2017). URL: https://population.un.org/wpp/Download/Standard/Population/(дата звернення: 23.04.2019).
4. Шекера О.Г. Демографічна ситуація у світі та в Україні / Науково-практичний журнал «Здоров'я суспільства», 2014.
5. Власенко Н.С., Макарова О.В., Пирожков С.І., та інші. Комплексний демографічний прогноз України на період до 2050 р. / за ред. член-кореспондент НАНУ, д.е.н., проф. Е.М. Лібанової. К.: Український центр соціальних реформ, 2006. 138 с.
6. Цвігун І.А. Демографічна безпека України та напрями її регулювання: монографія / Кам’янець-Подільський: Видавець ПП Зволейко Д.Г., 2013. 400 с.
7. Пальян З. О. Навчальний посібник Демографічна статистика / Київський Національний Економічний Університет України, 2003 р. 167 с.
8. Бідюк П.І., Коршевнюк Л.О., Проектування комп'ютерних інформаційних систем підтримки прийняття рішень: Навчальний посібник / Київ: ННК «Інститут прикладного системного аналізу»

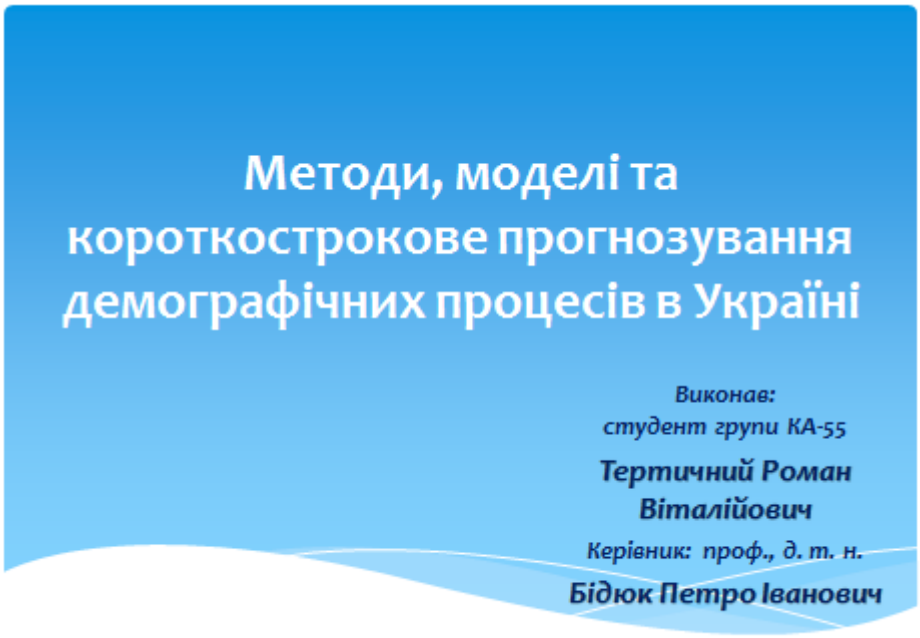
Національний технічний університет України «Київський політехнічний інститут», 2010. 340 с.

9. Бідюк П.І., Половцев О.В. Аналіз та моделювання економічних процесів перехідного періоду / К: ПЛАБ-75, 1999. 230 с.
10. Бідюк П.І. Системний підхід до побудови математичних моделей на основі часових рядів / Системні дослідження та інрормаційні технології, №3, 2002. l 14-131 с.
11. Бідюк П.І. Часові ряди: моделювання та прогнозування / Київ: ЕКМО, 2004.144 с.
12. Пашин В.П. Функционально-стоимостный анализ конструкторско-технологических решений. - К.: РДЭНТП «Знание» УССР, 1989. - 22с.
13. Пашін В.П. Оцінка конкурентоспроможності електронних пристроїв на стадії проектування. - К. Економічний вісник НТУУ „КПІ", 2006. - №3. с. 252-255.
14. Пашин В.П. Управление качеством изделий на основе функционально-стоимостного анализа. - К.: «Технология и организация производства», 1989. - №1. с. 17-19.
15. Пашін В. П. Методичні вказівки до виконання економіко-організаційного розділу дипломних проектів (робіт) бакалаврів і спеціалістів для студентів Інституту прикладного системного аналізу: Навч. посібник / Пашін В. П., Романов В. В., Єгорова Н. В – К.: НТУУ “КПІ”, 2011. – 118 с.

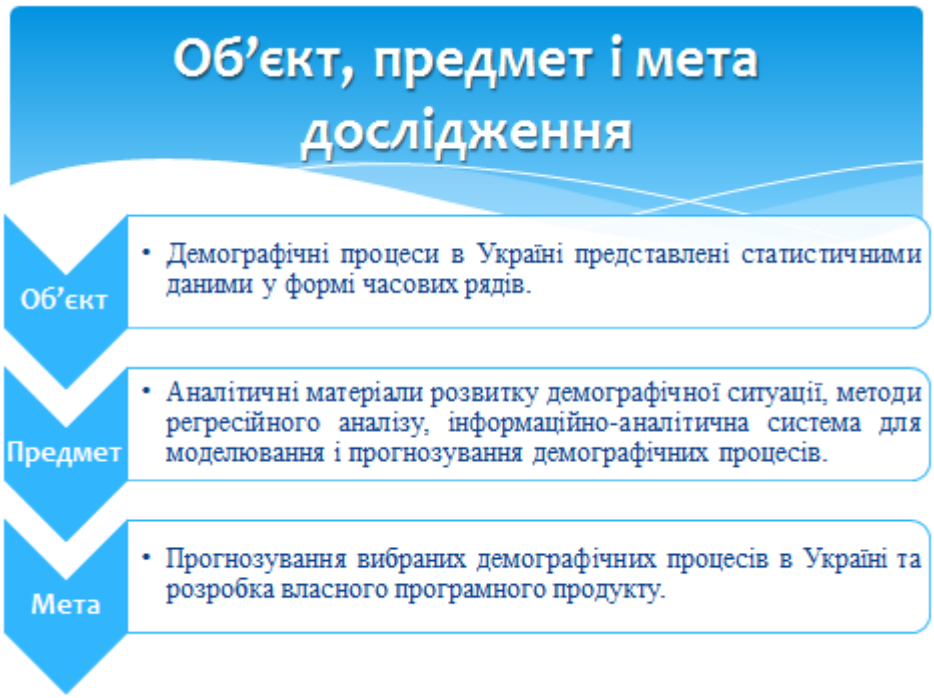
# ДОДАТОК А

Ілюстративний матеріал

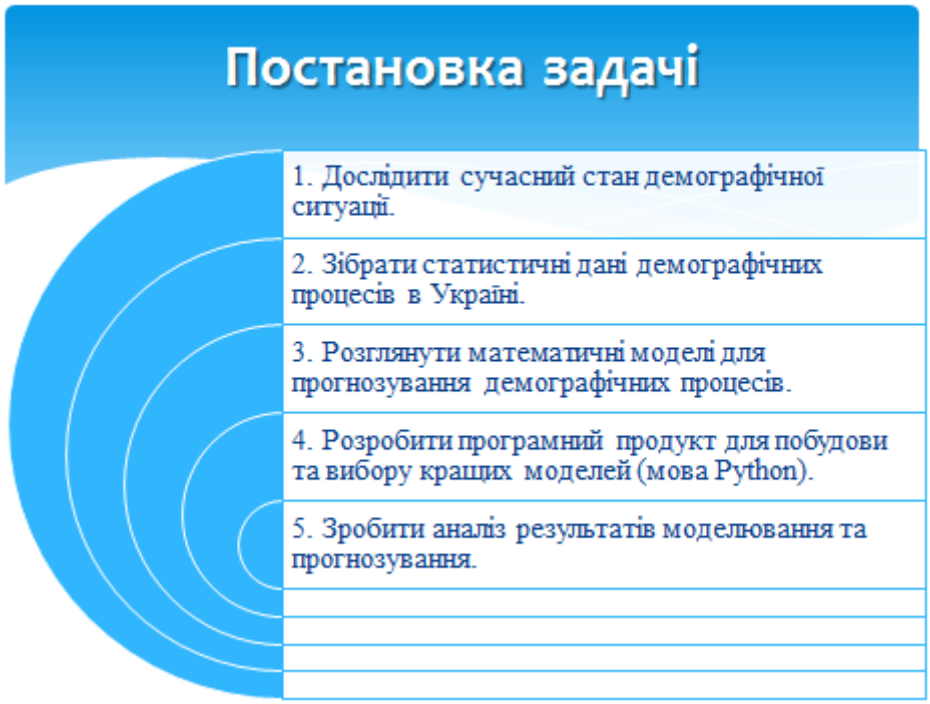
Слайд 1



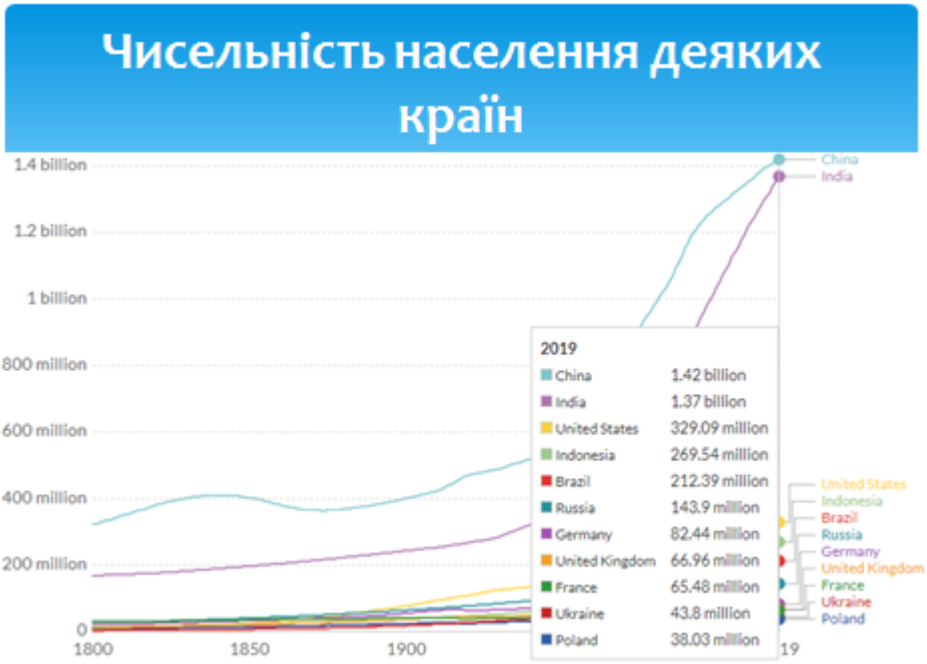

Слайд 2



Слайд 3

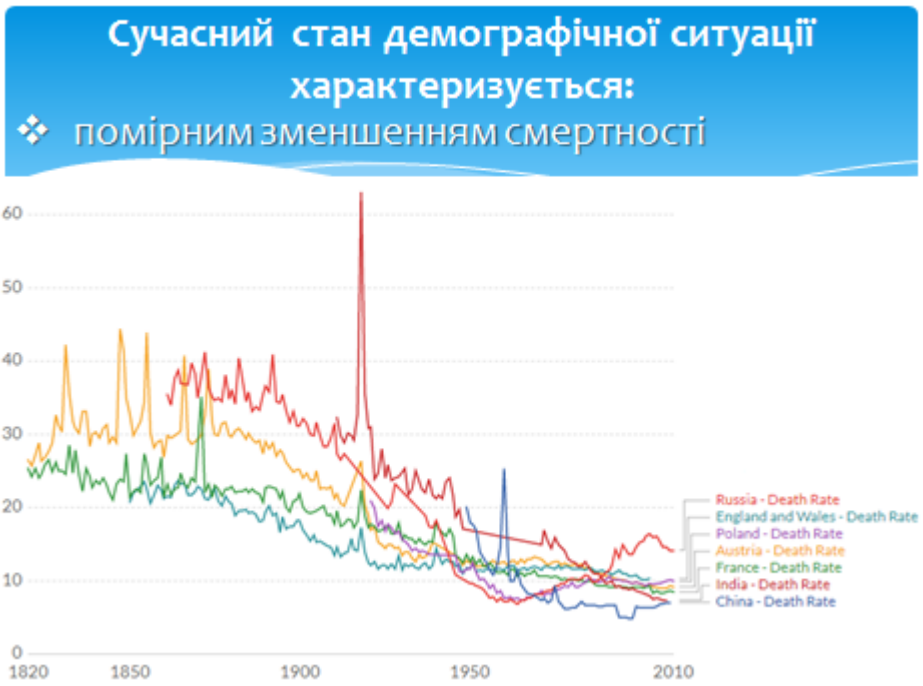



Слайд 4

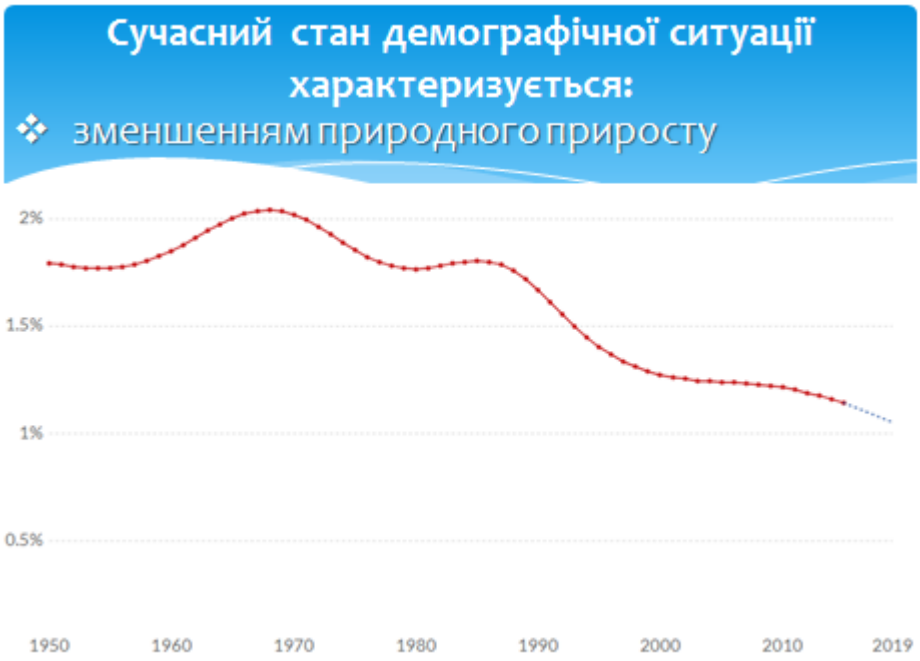


Слайд 5


Слайд 6

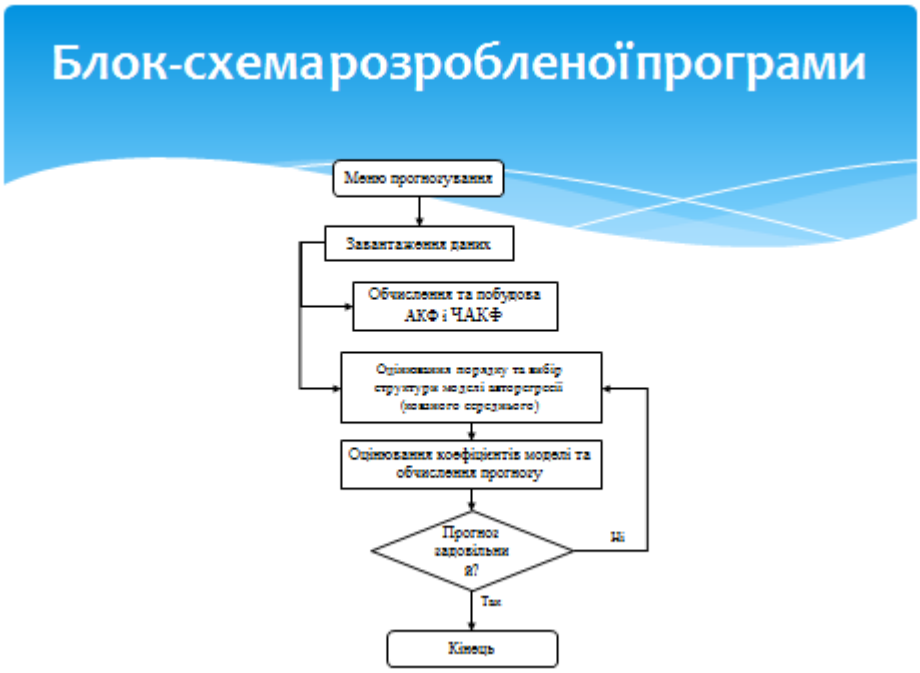


Слайд 7

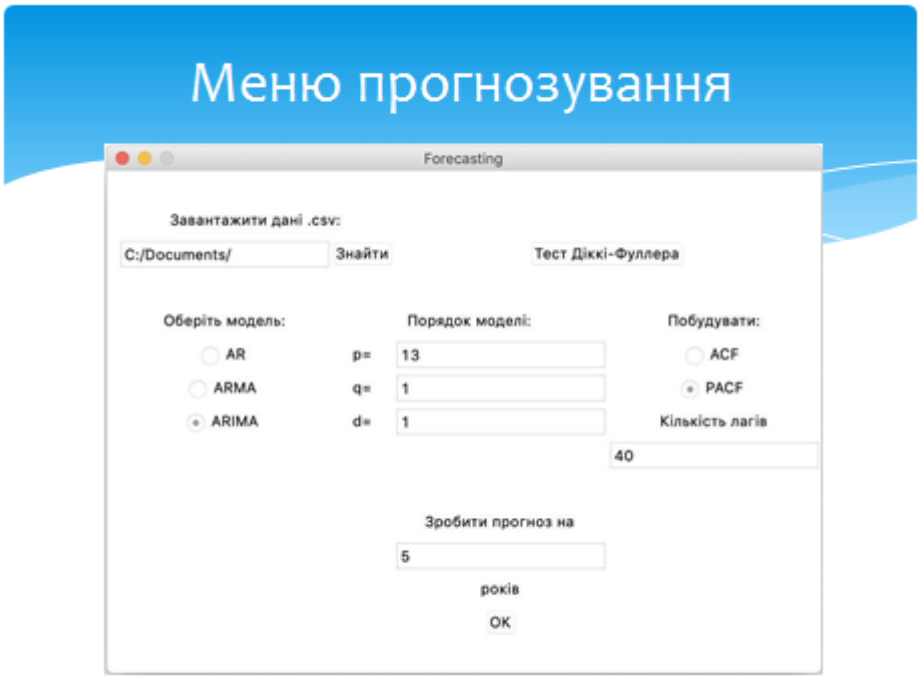



Слайд 8

Слайд 9




Слайд 10

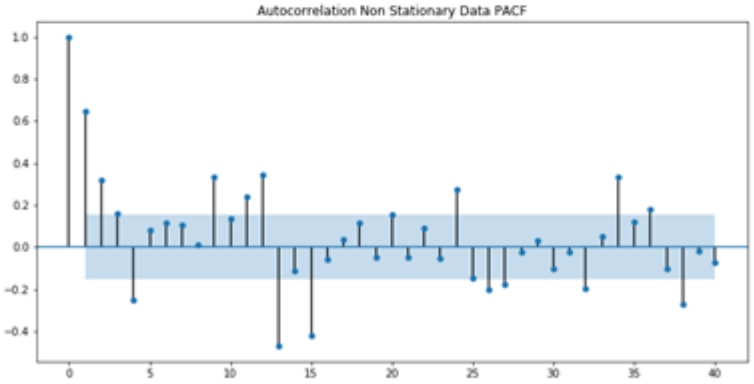


Слайд 11


Слайд 12

Слайд 13

## Приклади роботи програми для моделювання народжуваності: PACF




Слайд 14

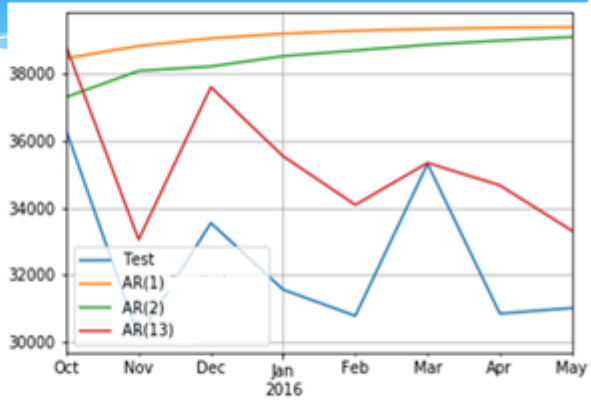
## Візуалізація моделей AR(1), AR(2), AR(13)

**AR(1):**

$$y(k) = 3.7381 + 0.621858\ y(k-1) + \varepsilon(k)$$

**AR(2):**

$$y(k) = 2.523 + 0.410139\ y(k-1) + 0.325851\ y(k-2) + \varepsilon(k)$$




**AR(13):**

$$y(k)=2.012 + 0.3734\ y(k-1) + 0.245937\ y(k-2) + 0.168274\ y(k-3) - 0.171265\ y(k-4) - 0.005166\ y(k-5) + 0.28109\ y(k-6) + 0.019134\ y(k-7) - 0.100341\ y(k-8) + 0.011192\ y(k-9) + 0.031772\ y(k-10) + 0.201880\ y(k-11) + 0.501992\ y(k-12) - 0.408821\ y(k-13) + \varepsilon(k)$$

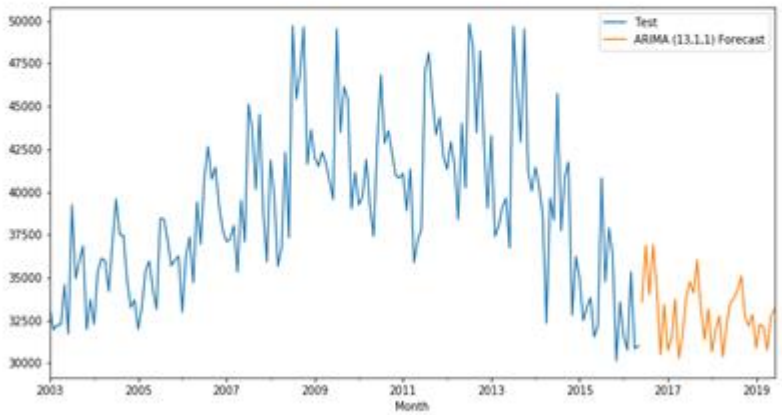
Слайд 15

## Висновки, отримані на основі побудованих моделей

| Тип моделі | Критерії адекватності | | | Оцінювання прогнозу | | |
|---|---|---|---|---|---|---|
| | $R^2$ | $\sum e^2(k)$ | $DW$ | СеКП | САПП | Theil |
| AR(1) | 0.395 | 1.112 | 2.398 | 0.079 | 0.631 | 0.0039 |
| AR(2) | 0.408 | 1.067 | 2.023 | 0.084 | 0.601 | 0.0037 |
| AR(13) | 0.607 | 0.889 | 1.859 | 0.078 | 0.571 | 0.0034 |
| ARMA(13,1) | 0.695 | 0.493 | 1.989 | 0.059 | 0.479 | 0.0029 |
| ARIMA(13,1,1) | 0.768 | 0.345 | 1.969 | 0.048 | 0.363 | 0.0023 |

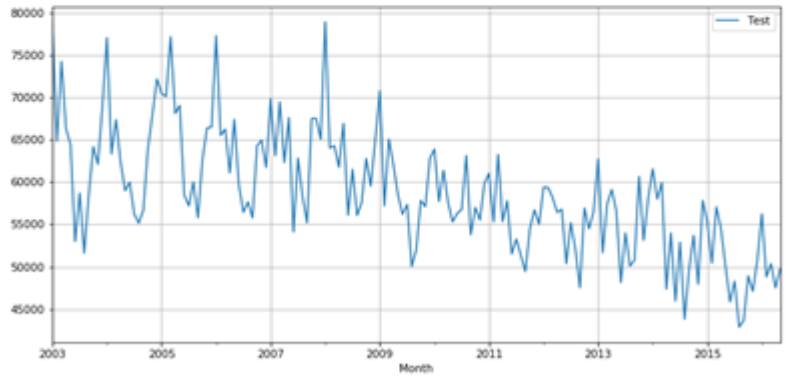
Слайд 16

## Прогноз народжуваності України на 3 роки



Слайд 17

## Реальний стан смертності за період 2003-2016 рр.




Слайд 18

## Висновки, отримані на основі побудованих моделей

| Тип моделі | Критерії адекватності | | | Оцінювання прогнозу | | |
|---|---|---|---|---|---|---|
| | $R^2$ | $\sum e^2(k)$ | $DW$ | СеКП | САПП | Theil |
| AR(1) | 0.397 | 1.298 | 2.495 | 0.090 | 0.668 | 0.0038 |
| AR(2) | 0.505 | 1.089 | 2.003 | 0.082 | 0.606 | 0.0038 |
| AR(12) | 0.759 | 0.501 | 1.937 | 0.061 | 0.459 | 0.0030 |
| ARIMA(12,1,1) | 0.768 | 0.498 | 1.945 | 0.055 | 0.419 | 0.0025 |

Слайд 19


Слайд 20

Слайд 21

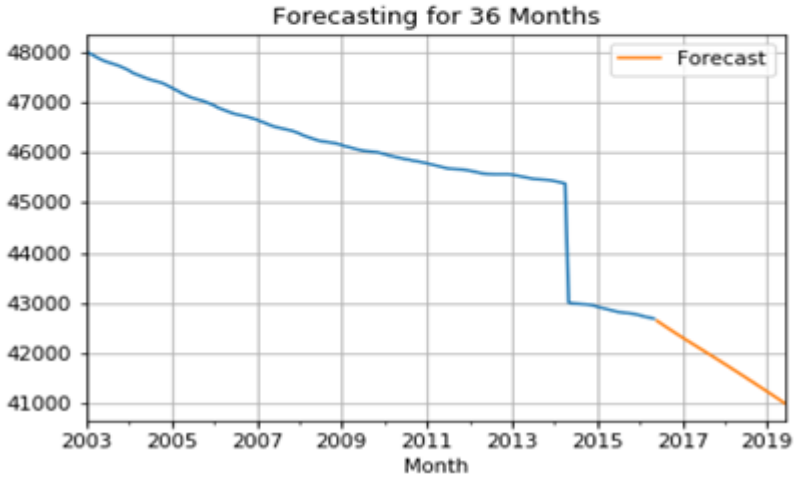
## Коректність побудови моделі AR(1)

Оскільки дані 2014 року вносять нелінійність в процес, розіб'ємо побудову моделі AR(1) на систему двох рівнянь, де межою поділу буде $k = 2014$:

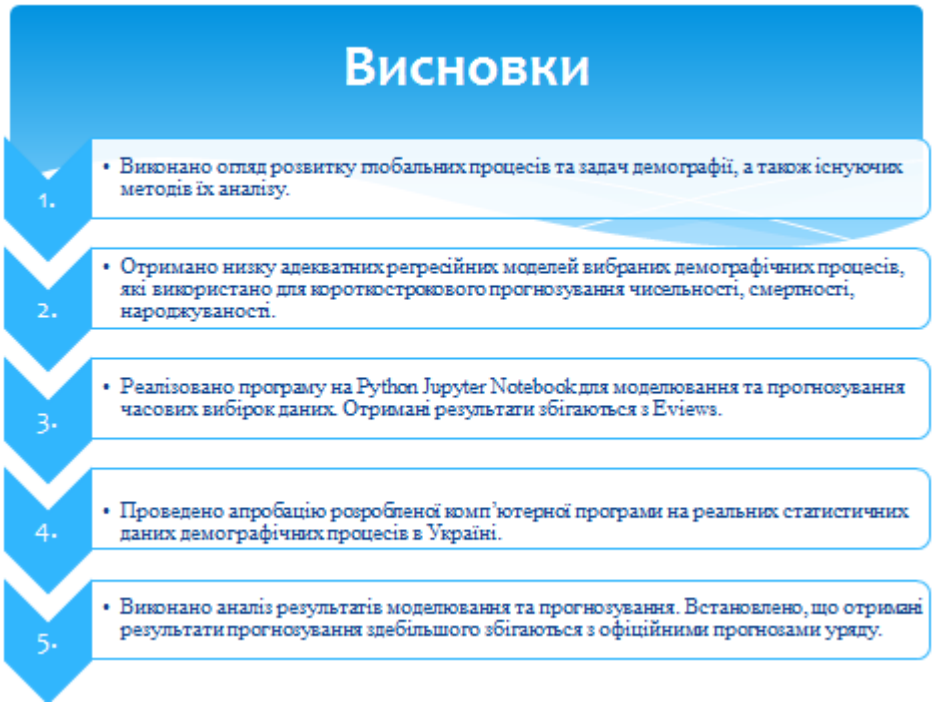
$$y(k) = \begin{cases} 0{,}0056 + 0{,}989712\ y(k-1) + \varepsilon(k), & k \leq 2014 \\ -0{,}0104 + 1{,}001535\ y(k-1) + \varepsilon(k), & k > 2014 \end{cases}$$

Слайд 22

## Прогноз чисельності населення України на 3 роки



Слайд 23




Слайд 24

Слайд 25

**Дякую за увагу!**

# ДОДАТОК Б

## Код програмного продукту

**ARmodel**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
#df =
pd.read_csv('https://raw.githubusercontent.com/Rom
aTertychnyi/DateSetDiplom/master/death.csv',index_
col='Month',parse_dates=True)
#df =
pd.read_csv('https://raw.githubusercontent.com/Rom
aTertychnyi/DateSetDiplom/master/born.csv',index_c
ol='Month',parse_dates=True)
df =
pd.read_csv('https://raw.githubusercontent.com/Rom
aTertychnyi/DateSetDiplom/master/populationAR(12
).csv',index_col='Month',parse_dates=True)
df.index.freq = None
df.head()
df.plot(figsize=(12,6),grid=True);
len(df)
train = df.iloc[:101]
test = df.iloc[101:]
print(train.shape, test.shape)
"""### Fit the model"""
from statsmodels.tsa.ar_model import AR,
ARResults
model = AR(train['Test'])
# Order 1 p=1 AR(1)
AR1fit =
model.fit(maxlag=1,method='cmle',trend='c',solver='l
bfgs')
# To know the order
AR1fit.k_ar
AR1fit.params
"""### Predict"""
start = len(train)
start
end=len(train) + len(test) - 1
end
pred1 = AR1fit.predict(start=start,end=end)
pred1 = pred1.rename('AR(1) Predictions')
test.plot(figsize=(12,6))
pred1.plot(legend=True);
# Order 2 p=2 AR(2)
AR2fit =
model.fit(maxlag=2,method='cmle',trend='c',solver='l
bfgs')
# Order
AR2fit.k_ar
# Parameters
AR2fit.params
pred2 = AR2fit.predict(start=start,end=end)
pred2 = pred2.rename('AR(2) Predictions')
pred2
test.plot(figsize=(12,6))
pred1.plot(legend=True)
pred2.plot(legend=True);
"""### Let Statsmodel choose the order for us"""
from statsmodels.tsa.ar_model import AR,
ARResults
model = AR(train['Test'])
ARfit = model.fit(ic='t-stat')
ARfit.k_ar # to know the right order
ARfit.params # to know all the parameters
pred11 = ARfit.predict(start=start,end=end)
pred11 = pred11.rename('AR(11) Predictions')
"""### Evaluate the model"""
from sklearn.metrics import mean_squared_error
labels = ['AR1','AR2','AR11']
preds = [pred1,pred2,pred11]
for i in range(3):
    error = mean_squared_error(test['Test'],preds[i])
    print('%s: Mean Squared Error =
%s'%(labels[i],error))
test.plot()
pred1.plot(legend=True)
pred2.plot(legend=True)
pred11.plot(legend=True)
plt.grid(True);
plt.savefig('AR_img.png');
# Forecast for Furture Values
model = AR(df['Test']) # Refit on the entire Dataset
ARfit = model.fit() # Refit on the entire Dataset
forecasted_values =
ARfit.predict(start=len(df),end=len(df)+36) #
Forecasting 3 year = 36 months
forecasted_values =
forecasted_values.rename('Forecast')
forecasted_values
# Plotting
df['Test'].plot(title='Forecasting for 36 Months')
forecasted_values.plot(legend=True)
plt.grid(True);
plt.savefig('AR_Forecast_img.png');
```

**ARIMAmodel**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
df1 =
pd.read_csv('https://raw.githubusercontent.com/Rom
aTertychnyi/DateSetDiplom/master/born.csv',index_c
ol='Month',parse_dates=True)
df1.index.freq = None
df1.head()
df1.plot(figsize=(12,6));
from statsmodels.tsa.seasonal import
seasonal_decompose
result = seasonal_decompose(df1['Test'],model='add')
from pylab import rcParams
rcParams['figure.figsize'] = 12,6
result.plot();
# Differencing once
from statsmodels.tsa.statespace.tools import diff
df1['Diff_1'] = diff(df1['Test'],k_diff=1)
df1['Diff_1'].plot(figsize=(12,6),title='Time Series
Differenced');
from statsmodels.tsa.stattools import adfuller
def adf_test(series,title=''):
    print(f'Augmented Dickey-Fuller Test: {title}')
    result = adfuller(series.dropna(),autolag='AIC')
    labels = ['ADF test statistic','p-value','# lags
used','# observations']
    out = pd.Series(result[0:4],index=labels)
    for key,val in result[4].items():
        out[f'critical value ({key})']=val
    print(out.to_string())
    if result[1] <= 0.05:
        print("Strong evidence against the null
hypothesis")
        print("Reject the null hypothesis")
        print("Data has no unit root and is stationary")
    else:
        print("Weak evidence against the null
hypothesis")
        print("Fail to reject the null hypothesis")
        print("Data has a unit root and is non-
stationary")
adf_test(df1['Test'],'Dickey-Fuller Test No Diff')
adf_test(df1['Diff_1'],'Dickey-Fuller Test Diff Once')
"""### 3. ACF PACF"""
from statsmodels.graphics.tsaplots import plot_acf
# just 40 lags is enough
plot_acf(df1['Test'],lags=40,title='Autocorrelation
Non Stationary Data');
plt.savefig('ARIMAforBorn_ACF_img.png');
from statsmodels.graphics.tsaplots import plot_pacf
# just 40 lags is enough
# shaded region is a 95 percent confidence interval
# Correlation values OUTSIDE of this confidence
interval are VERY HIGHLY LIKELY to be a
CORRELATION
plot_pacf(df1['Test'],lags=40,title='Autocorrelation
Non Stationary Data PACF');
plt.savefig('ARIMAforBorn_PACF_img.png');
len(df1)
train = df1.iloc[:153]
test = df1.iloc[153:]
"""### 5. ARIMA Model"""
from statsmodels.tsa.arima_model import ARIMA,
ARIMAResults
model = ARIMA(train['Test'],order=(13,1,1))
results = model.fit()
results.summary()
"""### 6. Predictions"""
start = len(train)
end = len(train) + len(test) - 1
# typ= 'levels' to return the differenced values to the
original units
preds =
results.predict(start=start,end=end,typ='levels').renam
e('ARIMA (13,1,1) Predictions')
preds
"""### 7. Plotting"""
test['Test'].plot(figsize=(12,6),legend=True)
preds.plot(legend=True);
train['Test'].plot(figsize=(12,6),legend=True)
test['Test'].plot(legend=True)
preds.plot(legend=True);
plt.savefig('ARIMA(13_1_1)Forecast+Test_img.png'
);
"""### 8. Evaluate the Model"""
from statsmodels.tools.eval_measures import rmse
error = rmse(test['Test'],preds)
error
test['Test'].mean()
(error/test['Test'].mean()) *100
"""### 9. Forecast for Future Data"""
# Refit with all the Data
model = ARIMA(df1['Test'],order=(13,1,1)) # Order
is chosen from Pyramid ARIMA
results = model.fit()
results.summary()
start = len(df1)
end = len(df1) + 36
# typ= 'levels' to return the differenced values to the
original units
forecasted_values =
results.predict(start=start,end=end,typ='levels').renam
e('ARIMA (13,1,1) Forecast')
df1['Test'].plot(figsize=(12,6),legend=True)
forecasted_values.plot(legend=True);
plt.savefig('ARIMA(13_1_1)Forecast_img.png');
```

**ARMAmodel**

```
mport pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
# Stationary
df1 =
pd.read_csv('https://raw.githubusercontent.com/Rom
aTertychnyi/DateSetDiplom/master/death.csv',index_
col='Month',parse_dates=True)
df1.index.freq = None
df1 = df1[:161] # Cogemos los 4 primeros meses, de
esta manera en mucho más estacionario para ARMA
df1.plot();
from statsmodels.tsa.arima_model import ARMA,
ARMAResults
df1['Test'].plot(figsize=(12,6),grid=True);
df1['Test'].rolling(window=15).mean().plot(figsize=(
16,6),grid=True,label='7 Day Window',legend=True);
"""### 1.1 Augmented Dickey-Fuller Test"""
from statsmodels.tsa.stattools import adfuller
def adf_test(series,title=''):
    print(f'Augmented Dickey-Fuller Test: {title}')
    result = adfuller(series.dropna(),autolag='AIC')
    labels = ['ADF test statistic','p-value','# lags
used','# observations']
    out = pd.Series(result[0:4],index=labels)
    for key,val in result[4].items():
        out[f'critical value ({key})']=val
    print(out.to_string())
    if result[1] <= 0.05:
        print("Strong evidence against the null
hypothesis")
        print("Reject the null hypothesis")
        print("Data has no unit root and is stationary")
    else:
        print("Weak evidence against the null
hypothesis")
        print("Fail to reject the null hypothesis")
        print("Data has a unit root and is non-
stationary")
adf_test(df1['Test'], 'Dickey-Fuller Test Births')
first_diff = df1['Test'] - df1['Test'].shift(1)
first_diff = first_diff.dropna(inplace = False)
adf_test(first_diff, 'Dickey-Fuller Test Births')
train = df1.iloc[:153]
test = df1.iloc[153:]
from statsmodels.tsa.arima_model import ARMA,
ARMAResults
#model = ARMA(train['Test'],order=(13,1)) #for
born
model = ARMA(train['Test'],order=(11,1)) #for death
results = model.fit()
results.summary()
"""#### Predict"""
start = len(train)
end = len(train) + len(test) - 1
preds =
results.predict(start=start,end=end).rename('ARMA
(11,1) Predictions')
"""### 1.5 Plotting the results"""
test['Test'].plot(figsize=(12,6))
preds.plot(legend=True);
plt.savefig('ARMA_Forecast_img.png');
test.mean()
preds.mean()
```